POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Improving Binary Classifier Performance Through an Informed Sampling Approach and Imputation
Auteur: Author:	Soroosh Ghorbani
Date:	2016
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ghorbani, S. (2016). Improving Binary Classifier Performance Through an Informed Sampling Approach and Imputation [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/2135/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/2135/
Directeurs de recherche: Advisors:	Michel C. Desmarais
Programme: Program:	Génie informatique

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

UNIVERSITÉ DE MONTRÉAL

IMPROVING BINARY CLASSIFIER PERFORMANCE THROUGH AN INFORMED SAMPLING APPROACH AND IMPUTATION

SOROOSH GHORBANI
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INFORMATIQUE)
AVRIL 2016

© Soroosh Ghorbani, 2016.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

IMPROVING BINARY CLASSIFIER PERFORMANCE THROUGH AN INFORMED SAMPLING APPROACH AND IMPUTATION

présentée par: GHORBANI Soroosh
en vue de l'obtention du diplôme de: Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de:

M. PESANT Gilles, Ph. D., président
M. DESMARAIS Michel C., Ph. D., membre et directeur de recherche
M. ANTONIOL Giuliano, Ph. D., membre
M. BOUGUESSA Mohamed, Ph. D., membre externe

DEDICATION

To my beloved parents
Ahmad and Farasat
who introduced me the joy of reading from birth,
to my loving spouse
Hanieh
for her patience, faith and unconditional love
and to the memory of a great friend
Dr. Yoosef Ramezani ...

ACKNOWLEDGEMENTS

It is a genuine pleasure, indeed honor, to extend my deepest gratitude to my supervisor, Prof. Michel C. Desmarais. His timely and insightful comments, meticulous scrutiny, enthusiasm and dynamism added considerably to my research experience and made an important part of this dissertation. I owe you my eternal gratitude for all I have learned from you Michel, Thank you!

I would like to thank all my lab colleagues and friends for collaborations, discussions, suggestions, and all the activities we did together.

On the personal side, I am very grateful to my family for their unflagging love and support throughout my life. My father and mother and also my wife, Hanieh have been, and will always be the source of strength and inspiration.

I am particularly thankful to my wife for her constant encouragement throughout my research period and for understanding all the late nights and weekends that went into finishing this thesis.

RÉSUMÉ

Au cours des deux dernières décennies, des progrès importants dans le domaine de l'apprentissage automatique ont été réalisés grâce à des techniques d'échantillonnage. Relevons par exemple le renforcement (boosting), une technique qui assigne des poids aux observations pour améliorer l'entraînement du modèle, ainsi que la technique d'apprentissage actif qui utilise des données non étiquetées partielles pour décider dynamiquement quels cas sont les plus pertinents à demander à un oracle d'étiqueter.

Cette thèse s'inscrit dans ces recherches et présente une nouvelle technique d'échantillonnage qui utilise l'entropie des données pour guider l'échantillonnage, un processus que nous appelons l'échantillonnage informé. L'idée centrale est que la fiabilité de l'estimation des paramètres d'un modèle peut dépendre de l'entropie des variables. Donc, l'adaptation du taux d'échantillonnage de variables basée sur leur entropie peut conduire à de meilleures estimations des paramètres.

Dans une série d'articles, nous étudions cette hypothèse pour trois modèles de classification, notamment Régression Logistique (LR), le modèle bayes naïf (NB) et le modèle d'arbre bayes naif (TAN—Tree Augmented Naive Bayes), en prenant une tâche de classification binaire avec une fonction d'erreur 0-1. Les résultats démontrent que l'échantillonnage d'entropie élevée (taux d'échantillonnage plus élevé pour les variables d'entropie élevée) améliore systématiquement les performances de prédiction du classificateur TAN. Toutefois, pour les classificateurs NB et LR, les résultats ne sont pas concluants. Des améliorations sont obtenues pour seulement la moitié des 11 ensembles de données utilisées et souvent les améliorations proviennent de l'échantillonnage à entropie élevée, rarement de l'échantillonnage à entropie faible.

Cette première expérience est reproduite dans une deuxième étude, cette fois en utilisant un contexte plus réaliste où l'entropie des variables est inconnue à priori, mais plutôt estimée avec des données initiales et où l'échantillonnage est ajusté à la volée avec les nouvelles estimation de l'entropie.

Les résultats démontrent qu'avec l'utilisation d'un ensemble de données initial de 1% du nombre total des exemplaires, qui variait de quelques centaines à environ 1000, les gains obtenus de l'étude précédente persistent pour le modèle TAN avec une amélioration moyenne de 13% dans la réduction l'erreur quadratique. Pour la même taille des semences, des améliorations ont également été obtenues pour le classificateur naïf bayésien par un facteur de 8% de l'entropie faible au lieu d'échantillonnage d'entropie élevée.

L'échantillonnage informé implique nécessairement des valeurs manquantes, et de nombreux classificateurs nécessitent soit l'imputation des valeurs manquantes, ou peuvent être améliorés par imputation. Par conséquent, l'imputation et l'échantillonnage informatif sont susceptibles d'être combinés dans la pratique. La question évidente est de savoir si les gains obtenus de chacun sont additifs ou s'ils se rapportent d'une manière plus complexe. Nous étudions dans un premier temps comment les méthodes d'imputation affectent la performance des classificateurs puis si la combinaison de techniques d'imputation avec l'échantillonnage informé apporte des gains qui se cumulent.

Le gain de méthodes d'imputation sont d'abord étudiés isolément avec une analyse comparative de la performance de certains nouveaux algorithmes et d'autres algorithmes d'imputation bien connus avec l'objectif de déterminer dans quelle mesure le motif des améliorations est stable dans les classificateurs pour la classification binaire. Ici encore, les résultats montrent que les améliorations obtenues par des techniques d'imputation peuvent varier considérablement par modèle et aussi par taux de valeur manquante. Nous étudions également les améliorations le long d'une autre dimension qui est de savoir si le taux d'échantillonnage par enregistrement est stable ou varie. Des différences mineures, mais statistiquement significatives sont observées dans les résultats, montrant que cette dimension peut également affecter les performances du classificateur.

Dans une dernière étude, nous étudions empiriquement si les gains obtenus de l'échantillonnage informé et de l'imputation sont additifs, ou s'ils se combinent d'une manière plus complexe. Les résultats montrent que les gains individuels de l'échantillonnage informé et d'imputation sont du même ordre de grandeur, mais en général, ils ne sont pas une simple somme des améliorations individuelles. Il faut noter aussi que, malgré les résultats encourageants pour certaines combinaisons d'échantillonnage informées et des algorithmes d'imputation, une analyse détaillée des résultats de l'ensemble de données individuelles révèle que ces combinaisons apportent rarement des performances supérieures aux algorithmes d'imputation ou à l'échantillonnage informé individuellement.

Les résultats de nos études fournissent une démonstration de l'efficacité de l'échantillonnage informé pour améliorer les performances de classification binaire pour le modèle TAN, mais les résultats sont plus mitigés pour NB et LR. En outre, l'échantillonnage à entropie élevée se révèle être le régime le plus bénéfique.

ABSTRACT

In the last two decades or so, some of the substantial advances in machine learning relate to sampling techniques. For example, boosting uses weighted sampling to improve model training, and active learning uses unlabeled data gathered so far to decide what are the most relevant data points to ask an oracle to label. This thesis introduces a novel sampling technique that uses features entropy to guide the sampling, a process we call informed sampling. The central idea is that the reliability of model parameter learning may be more sensitive to variables that have low, or high entropy. Therefore, adapting the sampling rate of variables based on their entropy may lead to better parameter estimates.

In a series of papers, we first test this hypothesis for three classifier models, Logistic regression (LR), Naive Bayes (NB), and Tree Augmented Naive Bayes (TAN), and over a binary classification task with a 0-1 loss function. The results show that the high-entropy sampling (higher sampling rate for high entropy variables) systematically improves the prediction performance of the TAN classifier. However, for the NB and LR classifiers, the picture is more blurry. Improvements are obtained for only half of the 11 datasets used, and often the improvements come from high-entropy sampling, seldom from low-entropy sampling. This first experiment is replicated in a second study, this time using a more realistic context where the entropy of variables is unknown a priori, but instead is estimated with seed data and adjusted on the fly. Results showed that using a seed dataset of 1% of the total number of instances, which ranged from a few hundreds to around 1000, the improvements obtained from the former study hold for TAN with an average improvement of 13% in RMSE reduction. For the same seed size improvements were also obtained for the Naive Bayes classifier by a factor of 8% from low instead of high entropy sampling. Also, the pattern of improvements for LR was almost the same as obtained from the former study.

Notwithstanding that classifier improvements can be obtained through informed sampling, but that the pattern of improvements varies across the informed sampling approach and the classifier model, we further investigate how the imputation methods affect this pattern. This question is of high importance because informed sampling necessarily implies missing values, and many classifiers either require the imputation of missing values, or can be improved by imputation. Therefore imputation and informative sampling are likely to be combined in practice. The obvious question is whether the gains obtained from each are additive or if they relate in a more complex manner.

The gain from imputation methods are first studied in isolation with a comparative analysis

of the performance of some new and some well known imputation algorithms, with the objective of determining to which extent the pattern of improvements is stable across classifiers for the binary classification and 0-1 loss function. Here too, results show that patterns of improvement of imputation algorithms can vary substantially per model and also per missing value rate. We also investigate the improvements along a different dimension which is whether the rate of sampling per record is stable or varies. Minor, but statistically significant differences are observed in the results, showing that this dimension can also affect classifier performance.

In a final paper, first the levels of improvement from informed sampling are compared with those from a number of imputation techniques. Next, we empirically investigate whether the gains obtained from sampling and imputation are additive, or they combine in a more complex manner. The results show that the individual gains from informed sampling and imputation are within the same range and that combining high-entropy informed sampling with imputation brings significant gains to the classifiers' performance, but generally, not as a simple sum of the individual improvements. It is also noteworthy that despite the encouraging results for some combinations of informed sampling and imputation algorithms, detailed analysis of individual dataset results reveals that these combinations rarely bring classification performance above the top imputation algorithms or informed sampling by themselves.

The results of our studies provide evidence of the effectiveness of informed sampling to improve the binary classification performance of the TAN model. Also, high-entropy sampling is shown to be the most preferable scheme to be conducted. This for example, in the context of Computerized Adaptive Testing, can be translated to favoring the highly uncertain questions (items of average difficulty). Variable number of items administered is another factor that should be taken into account when imputation is involved.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

RÉSUMÉ v

ABSTRACT vii

TABLE OF CONTENTS ix

LIST OF TABLES xii

LIST OF FIGURES xiv

LIST OF ACRONYMS AND ABBREVIATIONS xvii

LIST OF APPENDICES xviii

CHAPTER 1 INTRODUCTION 1
- 1.1 Overview and Motivations 1
- 1.2 Research Questions and Objectives 1
- 1.3 Summary of the Contributions 3
- 1.4 Organization of the Dissertation 3

CHAPTER 2 RELATED WORK 5
- 2.1 Planned Missing Data Designs 5
  - 2.1.1 Multiple Matrix Sampling 5
  - 2.1.2 Three-Form Design (And Variations) 6
  - 2.1.3 Growth-Curve Planned Missing 7
  - 2.1.4 Monotonic Sample Reduction 8
- 2.2 Active Learning 9
  - 2.2.1 Scenarios for Active Learning 10
- 2.3 Dealing with Missing Values in Classification Tasks 13
  - 2.3.1 Case Deletion 14
  - 2.3.2 Imputation 14
  - 2.3.3 Some Machine Learning Approaches 14

CHAPTER 3 SELECTIVE SAMPLING DESIGNS TO IMPROVE THE PERFORMANCE OF CLASSIFICATION METHODS . . . 15
3.1 Chapter Overview . . . 15
3.2 Entropy . . . 16
3.3 Models . . . 16
3.3.1 Naive Bayes . . . 16
3.3.2 Logistic Regression . . . 17
3.3.3 Tree Augmented Naive Bayes (TAN) . . . 17
3.4 Experimental Methodology . . . 18
3.4.1 Entropy-based heuristic for Selective Sampling . . . 18
3.4.2 The Process . . . 19
3.4.3 Datasets . . . 19
3.5 Results . . . 20
3.6 Conclusion . . . 23

CHAPTER 4 AN ADAPTIVE SAMPLING ALGORITHM TO IMPROVE THE PERFORMANCE OF CLASSIFICATION MODELS . . . 24
4.1 Chapter Overview . . . 24
4.2 Adaptive Sampling . . . 24
4.3 Methodology . . . 25
4.3.1 Adaptive Sampling and Seed Data . . . 25
4.3.2 Non-adaptive Selective Sampling . . . 26
4.3.3 Simulation Process . . . 26
4.4 Results . . . 26
4.5 Conclusion . . . 29

CHAPTER 5 PERFORMANCE COMPARISON OF RECENT IMPUTATION METHODS FOR CLASSIFICATION TASKS OVER BINARY DATA . . . 31
5.1 Chapter Overview . . . 31
5.2 Related work . . . 33
5.3 Description of Investigated Approaches . . . 35
5.3.1 The Nature of Missing Data . . . 35
5.3.2 Imputation Methods . . . 36
5.3.3 Classifier Models . . . 39
5.4 Experimental Methodology . . . 40
5.4.1 Datasets . . . 40
5.4.2 Experimental Setup . . . 40

5.5 Results 42
5.6 Conclusion 46

CHAPTER 6 IMPACT OF FIXED VS. VARIABLE SAMPLING RATE PER RECORD OVER THE PERFORMANCE OF IMPUTATION METHODS IN CLASSIFICATION TASKS 49
6.1 Chapter Overview 49
6.2 Description of the Imputation Methods and the Classifiers 50
6.3 Experimental Methodology 50
6.3.1 Rate and Distribution of Missing Values 51
6.3.2 Experimental Setup 51
6.4 Results 52
6.5 Conclusion 57

CHAPTER 7 INFORMED SAMPLING AND IMPUTATION METHODS IN BINARY CLASSIFICATION TASKS 59
7.1 Chapter Overview 59
7.2 Methodology 59
7.2.1 Experimental Setup 60
7.3 Results 62
7.3.1 First Experiment 62
7.3.2 Second Experiment 64
7.3.3 Detailed results 67
7.4 Conclusion 68

CHAPTER 8 CONCLUSION AND RESEARCH PERSPECTIVES 73
8.1 Conclusion 73
8.2 Limitations and Threats to Validity 76
8.2.1 Generalizability 76
8.3 Future Work 77

REFERENCES 79

APPENDICES 84

LIST OF TABLES

Table 2.1 Missing data pattern for a Three-Form design 7

Table 2.2 Missing data patterns for all combinations of one or two time points missing with 250 Subjects. Adapted from (Palmer and Royall, 2010) . 8

Table 2.3 An Example of Monotonic Sample Reduction 9

Table 3.1 Datasets - The Mean, Minimum and Maximum of the attribute entropies have been listed 20

Table 3.2 Performance comparison for the different techniques under the different schemes of sampling for Ketoprostaglandin-f1-alpha dataset (ICI-Incorrectly Classified Items and RMSE-Root-Means-Squared-Error) . 21

Table 3.3 Results of running a paired t-test on the obtained results of 100 folds based on average Pct. ICI 22

Table 3.4 Results of running a paired t-test on the obtained results of 100 folds based on average RMSE 22

Table 3.5 Percent of datasets on which Selective Sampling classification performance results are better ($p<0.05$) than Scheme 1 22

Table 3.6 Percent of datasets on which Scheme 1 classification performance results are better ($p<0.05$) than Scheme 2, Scheme 3 or both of them 23

Table 4.1 Performance comparison for the different techniques under the different schemes of sampling for Brain Chemistry dataset (ICI-Incorrectly Classified Items and RMSE-Root-Means-Squared-Error) where seed dataset size=8 27

Table 4.2 RMSE difference between scheme 1 and the two other schemes for Brain Chemistry dataset. Student-t test is based on 100 random sample simulations 28

Table 4.3 Number of datasets which show significant greater error (ARMSE) for each technique, under different sampling schemes, over 11 different datasets, and for different seed dataset sizes 29

Table 5.1 Imputation Methods Used in This Study 36

Table 5.2 Classifiers Considered in the Study. 39

Table 5.3 Datasets at a Glance. 40

Table 5.4 Percent of all the datasets on which applying the method improves the classification accuracy of the classifiers. 44

Table 7.1 Average F-Scores of NB classifier under different imputation/informed sampling methods over each dataset (Averaged over 100 Runs) 68
Table 7.2 Average F-Scores of LR classifier under different imputation/informed sampling methods over each dataset (Averaged over 100 Runs) 69
Table 7.3 Average F-Scores of TAN classifier under different imputation/informed sampling methods over each dataset (Averaged over 100 Runs) 69
Table A.1 Datasets on which the mean accuracies of the classifiers (over 10 runs and 6 missing rates) on imputed and non-imputed data are significantly different at $p < 0.05$ . . 84
Table A.2 Datasets on which the mean accuracies of the classifiers (over 10 runs) on imputed and non-imputed data are significantly different at $p < 0.05$. 85
Table A.3 Datasets on which none of the imputation methods improves the classification accuracy. 86

LIST OF FIGURES

Figure 1.1 The Thesis at a Glance. LE.S and HE.S are respectively low and high entropy sampling schemes, whereas U.S is uniform sampling as explained in Chapter 3. The terms single and multiple imputation are introduced in Chapter 5 and more information on Fixed vs. Variable rate sampling is given in Chapter 6 2

Figure 2.1 Multiple Matrix Sampling: dividing the interview questionnaire into sections of questions (shown in green) and then administering these sections to sub-samples of the main sample 6

Figure 2.2 Query Synthesis. Source: (Settles, 2012) 10

Figure 2.3 A handwriting recognition problem for which Query Synthesis works poorly when a human oracle is used. Source: (Settles, 2012) 11

Figure 2.4 Stream-Based Selective Sampling. Source: (Settles, 2012) 12

Figure 2.5 Pool-Based Sampling. Source: (Settles, 2012) 13

Figure 3.1 The Binary Entropy Function (MacKay, 2003) 16

Figure 3.2 a) Naive Bayes Classifier Structure and b) TAN Classifier Structure . 18

Figure 3.3 Sampling probability distribution used for the schemes 2 and 3 19

Figure 3.4 Ketoprostaglandin-f1-alpha Dataset 21

Figure 4.1 Adaptive Sampling Algorithm 25

Figure 4.2 Brain Chemistry Dataset 27

Figure 5.1 Main Steps in Multiple Imputation (here, m is assumed to be 3). 38

Figure 5.2 The General Procedure of the Experiments. 41

Figure 5.3 Comparison between Different Imputation Methods (Averaged over the 4 classifiers on all the datasets). 42

Figure 5.4 Classification Improvement for the Logistic Regression (top left), Tree Augmented Naïve Bayes (top right), Naïve Bayes (bottom left), and SVM with RBF Kernel (bottom right). 45

Figure 5.5 The plots represent the distribution of paired differences between the classification accuracy using each of the imputation methods and directly applying the data with missing values for each classifier. 47

Figure 6.1 Experimental Procedure. 52

Figure 6.2 Average classification accuracy of the models under different imputation methods over fixed vs. variable sampling schemes (averaged over the 14 datasets and 100 runs). 53

Figure 6.3 Orange) percentage of datasets on which the accuracy of a given model under a given imputation method over the variable sampling scheme is significantly higher than its accuracy under the same imputation method over the fixed scheme. Blue) similarly, shows the same percentage for the fixed sampling scheme against the variable scheme (Based on one tailed Wilcoxon Signed-Rank test at $p < 0.05$). . . 54

Figure 6.4 Inconsistent patterns of imputation results over fixed vs. variable sampling for: A) LR over dataset D10 at missing rate 30% and B) TAN over dataset D13 at missing rate 10%. 56

Figure 6.5 Percentage of datasets for which imputation results in inconsistent patterns of effects over variable vs. fixed sampling for each classifier. 57

Figure 7.1 Comparison. Experimental Setup. LE.S and HE.S are respectively low and high entropy sampling schemes, whereas U.S is the uninformed, uniform sampling which is used for the imputation methods. The performance of the different classifiers over the resulting datasets is then compared 60

Figure 7.2 Combination Experimental Setup. The imputation methods are applied on the three samples created by HE.S, LE.S and U.S schemes. The performance of the different classifiers over the resulting datasets is then compared 61

Figure 7.3 The Average gain (loss) in error reduction using formula (7.2) over 14 datasets and 100 runs. High entropy sampling can provide the TAN model with a substantial performance improvement that is comparable to imputation methods (HD, MF and MIEM). However, no gain is obtained for NB, and for the LR model two imputation methods bring substantial improvements (HD and MIEM) 63

Figure 7.4 Results of Wilcoxon signed-rank test score improvements for individual datasets at $p < 0.05$. Orange bars show the percentage of datasets on which $F.Score_i > F.Score_{base}$ and the blue bars show the percentage of cases on which $F.Score_i < F.Score_{base}$ where $i \in \{HD, MF, MILR, MIEM, LE.S, HE.S\}$ 64

Figure 7.5 The average gain over imputation, $Gain_I$ (eq. 7.3), when the informed sampling and imputation are applied in tandem. For NB, high entropy sampling scheme with MILR results in an average $Gain_I$ of 38%. For LR, when HE.S scheme is coupled with MILR or MIEM an average $Gain_I$ of 18.5% or 1.1% can be obtained respectively. Finally, for TAN, HE.S with MF, MILR and MIEM brings the average $Gain_I$ of 8%, 35% and 2.5% respectively 65

Figure 7.6 Results of Wilcoxon signed-rank test score improvements for individual datasets at $p < 0.05$. Orange bars show the percentage of datasets on which a significant difference is found for $F.Score_{s,i} > F.Score_{U.S,i}$ and the blue bars show the percentage of cases on which $F.Score_{s,i} < F.Score_{U.S,i}$ where where $i \in \{HD, MF, MILR, MIEM\}$ and $s \in \{LE.S, HE.S\}$ 66
Figure 7.7 Individual dataset logit of F-scores for NB. Bars are shown relative to the baseline 70
Figure 7.8 Individual dataset logit of F-scores for LR. Bars are shown relative to the baseline 71
Figure 7.9 Individual dataset logit of F-scores for TAN. Bars are shown relative to the baseline 72

LIST OF ACRONYMS AND ABBREVIATIONS

AANN	Auto-Associative Neural Networks
ARMSE	Average Root Mean Square Error
CART	Classification And Regression Tree
CAT	Computerized Adaptive Testing
DF	Degree of Freedom
DF	Degrees of Freedom
EC	Event Covering
EM	Expectation Maximization
GRNN	Generalized Regression Neural Network
HD	Hot deck
HE.S	High Entropy Sampling
ICI	Incorrectly Classified Items
K-NN	K-Nearest Neighbors
LE.S	Low Entropy Sampling
LR	Logistic Regression
MAR	Missing at Random
MCAR	Missing Completely at Random
MF	missForest
MICE	Multivariate Imputation by Chained Equations
MIEM	Multiple Imputation Based on Expectation Maximization
MILR	Multiple Imputation Based on Logistic Regression
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
NB	Naive Bayes
NMAR	Not Missing at Random
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RMSE	Root Mean Square Error
SOM	Self-Organizing Map
SVM	Support Vector Machine
TAN	Tree Augmented Naive Bayes
U.S	Uniform Sampling

LIST OF APPENDICES

Appendix A Some More Analysis 84

CHAPTER 1 INTRODUCTION

1.1 Overview and Motivations

Contrary to the old adage that the best solution to missing data is not to have them (Graham, 2009), there are times when wisely managing or building missing data into the overall measurement design is the best use of limited resources. In some context, we can choose to allocate the observations differently among the variables during the data gathering phase, such as training a model to perform Computerized Adaptive Testing (CAT). For adaptive testing, we have the opportunity to decide upon a specific scheme of missing values for each item. We can decide which subset of questions we wish to administer to each examinee during the data gathering phase, leaving unanswered items as missing values. Recommender systems, where initial suggestions provide seed data to construct the user profile, is another example.

Given a fixed number of observations, we need to determine which variables are most critical (the most relevant information) and, potentially, ought to be allotted more observations and this leads us to optimizing the choice of missing values distribution through conducting an informed sampling process.

1.2 Research Questions and Objectives

We formalize our objectives through some concise research questions as follows:

- **RQ1**- Can a selective sampling approach based on uncertainty improve the performance of classifiers?
- **RQ2**- Can we guide the selective sampling approach on the fly, during the data gathering phase?

Informed selective sampling necessarily involves missing values and an obvious way to manage them is through imputation. Our study of informed sampling focuses on fixed sampling rate, since we would typically use informed sampling in a context where we have a fixed number of observations per record, or at least this is a plausible constraint in many contexts such as CAT or recommender systems as mentioned above. However the literature to date reports results of imputation techniques that use a variable rate of sampling (each record may contain

a different number of observations) for relatively small amounts of missing data. This leads to the following research question:

- **RQ3**- Can imputation techniques improve the prediction accuracy of classification tasks with a fixed rate of observations per record?

If the answer to the question above is positive, two more research questions can be stated as follows:

- **RQ4**- What is the impact of variable vs. fixed sampling rate per record on the performance of imputation methods in classification tasks?
- **RQ5**- Are the improvements from informed sampling and from imputation additive?

To address the questions above, this study revolves around improving the binary classification accuracy, as shown by figure 1.1. We explore the space created by the dimensions shown to investigate the effect of each dimension on classification performance, in particular to find out if the effect found in one dimension is independent of the others, or if they interact in a more complex way.

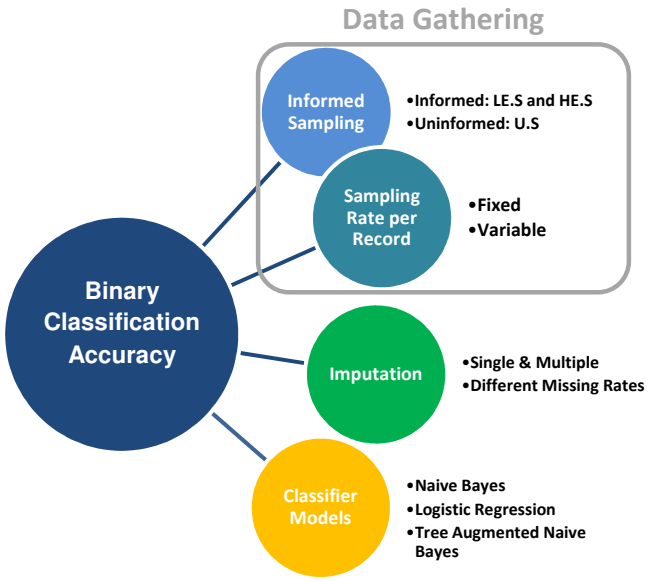

Figure 1.1 The Thesis at a Glance. LE.S and HE.S are respectively low and high entropy sampling schemes, whereas U.S is uniform sampling as explained in Chapter 3. The terms single and multiple imputation are introduced in Chapter 5 and more information on Fixed vs. Variable rate sampling is given in Chapter 6

1.3 Summary of the Contributions

The research presented in this dissertation has led to our main contributions listed as follows:

1. An entropy-based selective sampling design to improve the performance of classification models (Answering research question RQ1)
2. A demonstration that the gains from the proposed informed sampling approach are maintained when entropy is estimated on the fly (Answering research question RQ2)
3. An extensive assessment of the impact of imputation methods under different sampling schemes that resulted in the following demonstrations (Answering research questions RQ3 and RQ4):
   - Imputation can improve the performance of classifiers over data with fixed rate and high proportions of missing values per record
   - The choice of the rate of sampling per record between fixed and variable affects the performance of imputation methods
4. A comprehensive study on the potential additive effects of informed sampling and imputation methods on classification performance resulted in (Answering research questions RQ5) :
   - Showing that individual gains from informed sampling and imputation are within the same range
   - Demonstrating that the informed sampling often improves imputation algorithms performance, but in general the gains from the two are not additive

1.4 Organization of the Dissertation

The remainder of this dissertation provides the following content:

- **Chapter 2:** Presents the necessary background and introduces the related work relevant to this research.
- **Chapter 3:** Describes our proposed selective sampling approach which relies on features' uncertainty to improve the prediction performance of binary classification. The study investigates the effect of some entropy-based sampling schemes on the predictive performance of three different classification models namely, Naive Bayes, Logistic Regression and Tree Augmented Naive Bayes.

- **Chapter 4:** Presents another study in which we further explore the entropy-based heuristic to guide the sampling process on the fly. The same sampling schemes and classification models as mentioned above are used in this study.

- **Chapter 5:** Explains our study which evaluates the effect of two recently proposed imputation methods, namely missForest and Multiple Imputation based on Expectation-Maximization, and further investigates the effectiveness of two other imputation methods: Sequential Hot-deck and multiple imputation based on Logistic Regression on data with fixed rate of missing values per record. Their effect is assessed over the classification accuracy of four different models of classifiers with respect to varying amounts of missing data (i.e., between 5% and 50%).

- **Chapter 6:** Presents our study which has been done again in the context of imputation for classification purposes and investigates whether, and how the performance of different classifiers is affected by the two sampling schemes: fixed and variable rate of observations per record given equal number of observations in total. Three single imputation methods including Mean, missForest and Sequential Hot-deck, and 2 multiple imputation methods based on Logistic Regression and Expectation-Maximization are tested at 4 different amounts of missing data (ranging from 10% to 70%) with 3 different classifier models.

- **Chapter 7:** Demonstrates another study in which we investigate the possibility of combining the improvements (of classification prediction performance) from the informed sampling approach and different methods of missing values imputation. The same sampling schemes and the classification models explained in Chapter 3 are tested with the same imputation methods introduced in Chapter 5.

- **Chapter 8:** Presents the conclusion of this dissertation and outlines some directions of future research.

CHAPTER 2 RELATED WORK

In this chapter, we discuss relevant notable work to our research. Three main topics have been selected for the literature review. First, in section 2.1, the Planned Missing Data strategy and some of its different designs are briefly explained. Then in section 2.2, the paradigm of Active Learning is introduced and finally, in section 2.3, different approaches to deal with missing values in classification tasks are overviewed.

2.1 Planned Missing Data Designs

Selective Sampling is analogous to the notion of planned missing data designs used in psychometry and other domains. In planned missing data designs, participants are randomly assigned to conditions in which they do not respond to all items. Planned missing data is desirable when, for example:

- long assessments can reduce data quality, a situation that arises frequently when data is gathered from a human subject or some source for which a measurement has an effect on posterior measurements due to fatigue or boredom for example,
- data collection is time and cost intensive, and time/cost varies across attributes, in which case finding the optimal ratio of missing values over observation for each attribute is important.

Planned missing values were originally studied in the context of sampling theory and inferential statistics, but the issues are very similar to the ones found in the context of training a classifier, and in statistical learning algorithms in general. The data gathering phase of the learning algorithms may be faced with the same constraints as found in experimental design. Therefore, in this section, we briefly take a look at some planned missing data techniques applied in Cross-Sectional or Longitudinal studies.

2.1.1 Multiple Matrix Sampling

Multiple matrix sampling is one of the proposed methods to decrease the length of a given interview. This would involve dividing the interview questionnaire into sections of questions and then administering these sections to subsamples of the main sample, as figure 2.1 shows. It is assumed that the N items are a random sample from a population of items (just as M

participants are a random sample from a population) (Shoemaker, 1971). The most important gain in this procedure is that individuals are tested on only a portion of the test items in the total pool, and yet the parameters of the universe scores (mean of test score, variance of test score) can be accurately estimated. There are some considerations to be made in the applications of the multiple matrices sampling. These include: the number of subsets, the number of items per subset and the number of examinees administered each subset. These variables can be manipulated to create several multiple matrix sampling plans. The design to be adopted by the test developer will depend on the situations on the ground. These may be in the form of the number of available examinees, times available and the cost of materials. Several types of matrix sampling have been proposed (Anigbo, 2011).

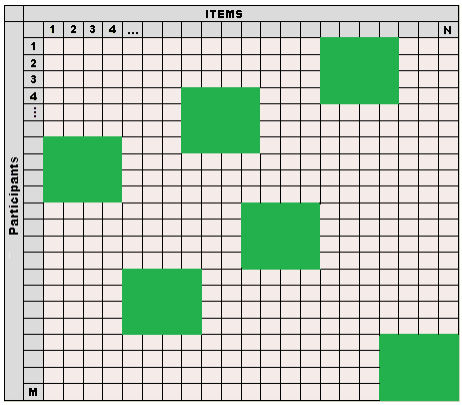

Figure 2.1 Multiple Matrix Sampling: dividing the interview questionnaire into sections of questions (shown in green) and then administering these sections to sub-samples of the main sample

2.1.2 Three-Form Design (And Variations)

In many research contexts the number of survey items can be excessive and overburden the respondent. The Three-Form design proposed by Graham et al. (1996, 2006) is a way to overcome this dilemma. This design creates three questionnaire forms, each of which is missing a different subset of items. The design divides the item pool into four sets (Common Set X, Set A, Set B, and Set C) and allocates these sets across three questionnaire forms, such that each form includes X and is missing A, B, or C. A layout of the basic design can be found in table 2.1. As an illustration of the 3-form design, suppose a researcher is interested in administering four questionnaires to a sample of sixth graders and each questionnaire has 30 items. The resulting questionnaire battery would include 120 items. The attention span required to respond to the entire questionnaire set would be too great for a sixth grader,

but the students could realistically respond to a shorter questionnaire with 90 items. Using the 3-form design, each sixth grader will only be required to respond to 90 items, but the researcher will be able to analyze the data based on the entire set of 120 items. The 3-form design is flexible and can be adapted to research needs. However, it requires careful planning as there are a number of important nuances in implementation (e.g., optimizing power by properly allocating questionnaires to the forms, constructing the forms in a way that allows for the estimation of interactive effect) (Baraldi and Enders, 2010). Also, Graham et al. (2001) describe variations of the 3-form design that can be applied to longitudinal studies. The basic idea is to split the sample into a number of random subgroups and impose planned missing data patterns, such that each subgroup misses a particular wave (or waves) of data. The idea of purposefully introducing missing data is often met with skepticism, but they show that planned missing data designs can be more powerful than complete-data design that use the same number of data points. Graham et al.'s results suggest that collecting complete data from N participants will actually yield less power than collecting incomplete data from a larger number of respondents (Baraldi and Enders, 2010).

Table 2.1 Missing data pattern for a Three-Form design

Form	Common Set X	Set A	Set B	Set C
1	25% of items	25% of items	25% of items	missing
2	25% of items	25% of items	missing	25% of items
3	25% of items	missing	25% of items	25% of items

2.1.3 Growth-Curve Planned Missing

By planning the missing data pattern across subjects, the surprising usefulness of using growth curve models has been demonstrated as one solution to the problem of respondent burden in ongoing longitudinal assessments (Graham et al., 2001). A growth curve is an empirical model of the evolution of a quantity over time. In growth-curve designs, the most important parameters are the growth parameters (e.g., estimate the steepness and the shape of the curve) and estimation precision depends heavily on the first and last time points. A planned missing design can take advantage of this by putting missingness in the middle. As an example, if 250 subjects were recruited for four waves of measurement, there could be five patterns of planned missing data. One set of 50 subjects would be assessed at all waves, another set of 50 subjects would miss the first follow-up but have data on subsequent waves, another 50 would be missing only at the second follow-up, and so forth. This scenario

would yield 20% missing data. In a planned missing scenario in which subjects miss two assessment points, missing data would be 40%. Table 2.2 shows the scenario for all possible combinations of missing data points for one and then two time points with the data missing by design (Palmer and Royall, 2010).

Furthermore, Graham et al showed that, despite identical costs, a planned missing design with 30% missing data had smaller standard errors (greater power to detect an effect) than the complete case design (Graham et al., 2001).

Table 2.2 Missing data patterns for all combinations of one or two time points missing with 250 Subjects. Adapted from (Palmer and Royall, 2010)

Subject	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 4**
All combinations of one missing time point (20% missing) (n=50 in each set)				
1	x	x	x	x
2	x	x	x	missing
3	x	x	missing	x
4	x	missing	x	x
5	missing	x	x	x
All combinations of two missing time points (42% missing) (n=36 in each set)				
1	x	x	x	x
2	x	x	missing	missing
3	x	missing	x	missing
4	missing	x	x	missing
5	x	missing	missing	x
6	missing	x	missing	x
7	missing	missing	x	x

2.1.4 Monotonic Sample Reduction

Monotonic sample reduction is sometimes used in large datasets (e.g., Early Childhood Longitudinal Study) to reduce costs. At each wave of measurement, a randomly-selected subgroup of the original sample is observed again. The remainder of the original participants does not need to be kept track of (treated as missing data) as table 2.3 shows. The main advantages of the approach are remarkable cost reduction and a lot of power to estimate effects that it yields at earlier waves.

Table 2.3 An Example of Monotonic Sample Reduction

Group	Time 1	Time 2	Time 3	Time 4	Time 5
1	x	x	x	x	x
2	x	x	x	x	missing
3	x	x	x	missing	missing
4	x	x	missing	missing	missing
5	x	missing	missing	missing	missing

2.2 Active Learning

Another close discipline to the informed sampling we study in this research is Active Learning. As a sub-field of machine learning, active learning is the study of computer systems that improve with experience and training. An active learning system develops and tests new hypotheses as part of a continuing, interactive learning process. It may ask queries, usually in the form of unlabeled data instances to be labeled by an oracle (e.g., a human annotator). For this reason, active learning is sometimes called "query learning" in the computational learning theory literature (Settles, 2012).

Having the learner ask questions or be more involved in its own training process can be very advantageous in many application contexts. For any supervised learning system to perform well, it must often be trained on hundreds (even thousands) of labeled instances. Sometimes these labels come at little or no cost through crowdsourcing, for example, the "spam" flag we mark on unwanted email messages, or the five-star rating we might give to movies on a website. Learning systems use these flags and ratings to better filter our junk email and suggest movies we might enjoy. In these examples we provide such labels for free, but we can find many other supervised learning tasks for which labeled instances are very difficult, time-consuming, or expensive to obtain. For one example, learning to classify documents or any other kind of media usually requires people to annotate each item with particular labels, such as relevant or not-relevant. Unlabeled instances are abundant from resources like the Internet, but annotating thousands of these instances can be long and tiresome and even redundant. In examples like this, data collection for traditional supervised learning systems can be very costly in terms of human effort and/or laboratory materials. If an active learning system is allowed to be part of the data collection process, where labeled data is scarce and unlabeled data abundant, it attempts to overcome the labeling bottleneck by adaptively requesting labels. In this way, the active learner aims to attain good generalization and achieve high accuracy using as few labeled instances as possible (Settles, 2010, 2012).

Therefore, active learning is most appropriate when the unlabeled data instances are numer-

ous, can be easily collected, and we anticipate having to label many of them to train an accurate system.

2.2.1 Scenarios for Active Learning

The learner may be able to ask queries in several different ways. The main scenarios that have been considered for active learning in the literature are as follows (Settles, 2012):

Query Synthesis

In this setting, for any unlabeled data instance in the input space as shown in figure[1] 2.2, the learner may request "label membership" including queries that the learner synthesizes de novo. The only assumption is that the learner has a definition of the input space (i.e., the feature dimensions and ranges) available to it.

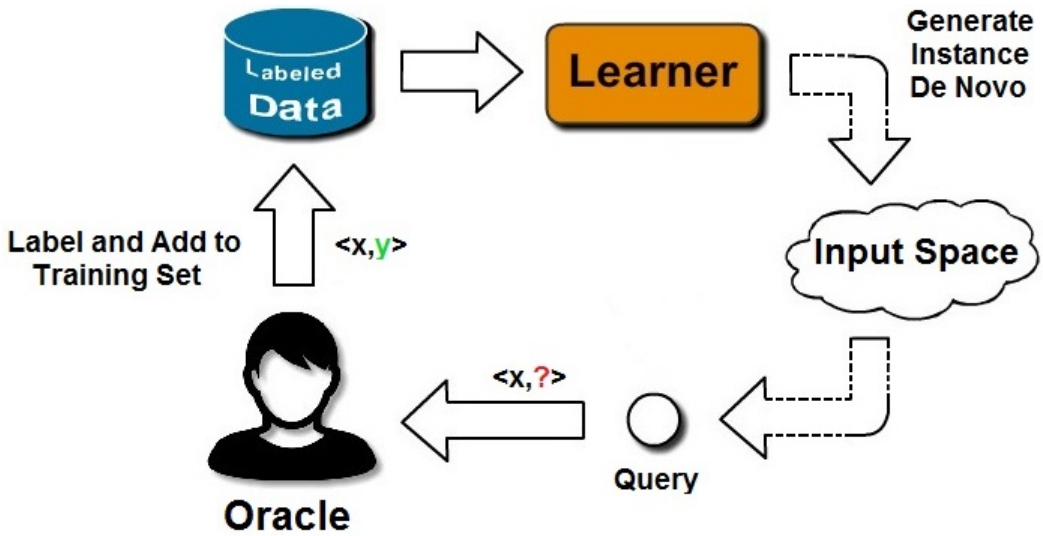

Figure 2.2 Query Synthesis. Source: (Settles, 2012)

As, for instance, shown in (King et al., 2004, 2009), query synthesis is reasonable for some real-world problems, but labeling such arbitrary instances can be awkward and sometimes troublesome. Figure 2.3 shows an unexpected problem Baum and Lang (1992) encountered when they tried to use membership query learning with human oracles to train a neural network to classify handwritten characters: many of the query images generated by the

[1]Figures 2.2 to 2.5 are reproduced with kind permission of the author and the publisher.

learner were merely artificial hybrid characters and not recognizable. As can be seen it is not clear what the image in the upper-right hand corner is, a 5, an 8, or a 9?

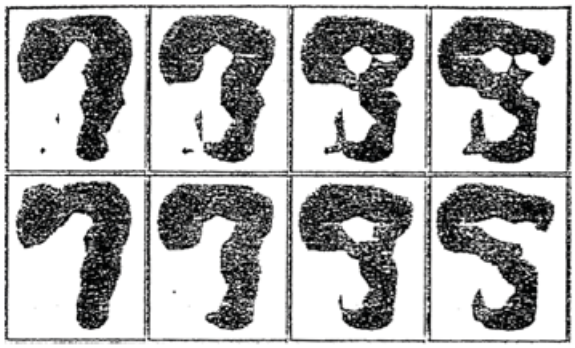

Figure 2.3 A handwriting recognition problem for which Query Synthesis works poorly when a human oracle is used. Source: (Settles, 2012)

Stream-Based Selective Sampling

In this scenario, it is assumed that obtaining an unlabeled instance is free (or inexpensive), so as illustrated by figure 2.4, it can first be sampled (typically one at a time) from the actual distribution, and then the learner can decide whether to request its label or discard it. This can be implemented in several ways. One approach can be to define a measure of utility (or information content), such that instances with higher utility are more likely to be queried (see e.g., (Dagan and Engelson, 1995)). More details on this approach and some others can be found in (Settles, 2012).

The stream-based active learning has been studied in several real-world tasks, including part-of speech tagging (Dagan and Engelson, 1995), sensor scheduling (Krishnamurthy, 2002), learning ranking functions for information retrieval (Yu, 2005) and word sense disambiguation in Japanese language (Fujii et al., 1998).

Pool-Based Sampling

As shown by figure 2.5, this scenario assumes that there is a small set of labeled data and a large pool of unlabeled data available. Queries are typically chosen in a greedy fashion from the pool, which is usually assumed to be non-changing, in accordance with a utility measure used to evaluate all instances in the pool (or a sub-sample of it in case it is very large).

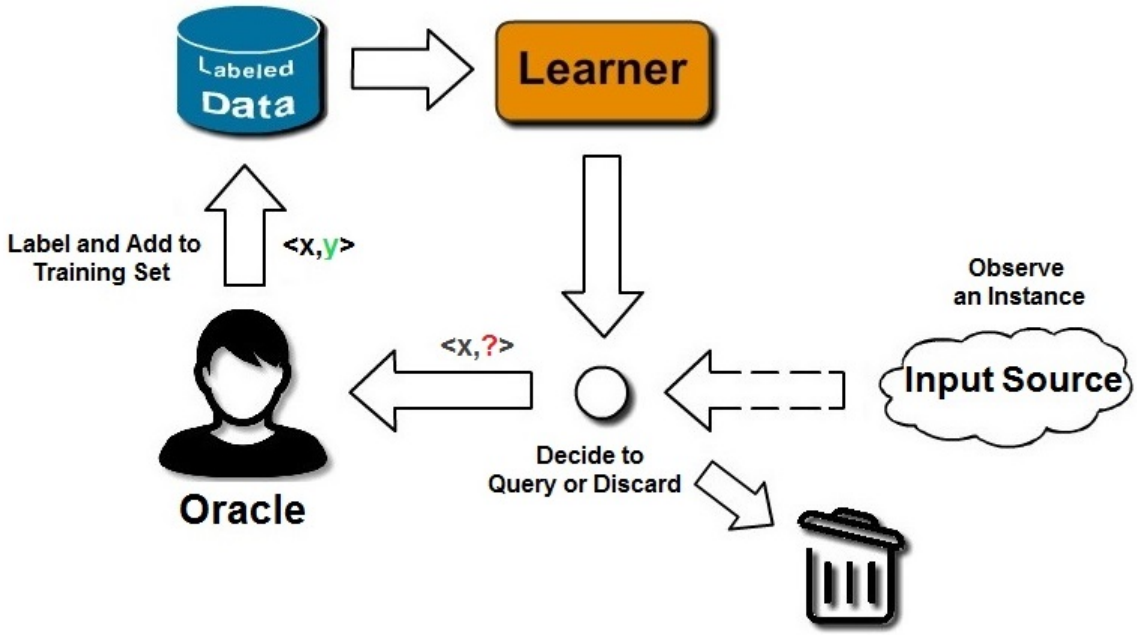

Figure 2.4 Stream-Based Selective Sampling. Source: (Settles, 2012)

This scenario has been studied for many real-world problem domains in machine learning, such as text classification (Lewis and Gale, 1994), information extraction (Settles and Craven, 2008), image classification and retrieval (Zhang and Chen, 2002) and speech recognition (Tur et al., 2005), to name only a few. In fact, compared to query synthesis and stream-based scenarios which are more common in the theoretical literature, pool-based sampling is the most popular scenario for applied research in active learning (Settles, 2012).

The main difference between stream-based and pool-based active learning strategies is that the former as mentioned earlier, obtains one instance at a time sequentially, and makes each query decision individually. On the contrary, pool-based active learning evaluates and ranks the entire U before selecting the best query (Settles, 2012).

Although, the informed sampling we study in this dissertation and active learning are similar in that they both affect the training set through collecting more informative data to be used in the learning process, the strategies they use to achieve this goal is completely different: informed sampling focuses on variables (by allotting more observations to the more relevant variables) and active learning concentrates on instances (by asking the oracle to label the most informative instances). The other differences between the two are:

- Informed sampling necessarily implies missing values, which is not the case for the

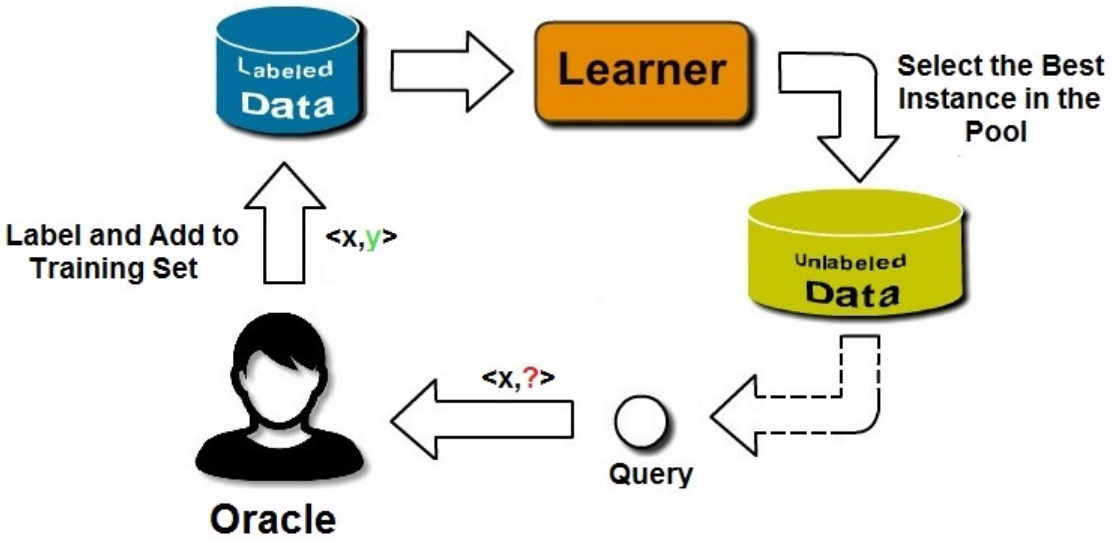

Figure 2.5 Pool-Based Sampling. Source: (Settles, 2012)

contexts in which active learning applies.

- Availability of abundant (unlabeled) data instances, as the main underlying assumption, is a necessity for active learning. Remembering the aforementioned CAT example, informed sampling does not have such prerequisite.

Because missing values are necessarily involved in informed sampling, we briefly look at the different strategies used in the literature to deal with missing values in the following section.

2.3 Dealing with Missing Values in Classification Tasks

Learning, inference, and prediction in the presence of missing values are pervasive problems in machine learning and statistical data analysis. The classification setting is particularly affected by the problem since many classifier models have no natural ability to deal with missing input features. Besides, missing values in either the training set or test set or in both sets affect the prediction accuracy of learned classifiers (Luengo et al., 2012). The appropriate way to handle incomplete input data depends in most cases on how data attributes became missing (three different mechanisms, which lead to the introduction of missing values are discussed in Chapter 5). Usually, the treatment of missing values in data mining can be handled in the following different ways (Luengo et al., 2012; García-Laencina et al., 2010).

2.3.1 Case Deletion

Case deletion is known as *complete case analysis*. It is available in all statistical packages and is the default method in many programs. This method discards all instances (cases) with missing values for at least one feature from the dataset. A variation of this method consists of determining the extent of missing data on each instance and feature, and delete the instances and/or features with high levels of missing data. Before deleting any feature, it is necessary to evaluate its relevance to the analysis. Relevant attributes should be kept even with a high degree of missing values.

Nevertheless, deletion methods are practical only when the data contain relatively small number of instances with missing values and when the analysis of the remaining data (complete instances in case deletion) will not lead to serious bias during the inference.

2.3.2 Imputation

The imputation of missing values is a class of procedures that aims to substitute the missing data with estimated values. Different approaches to the imputation of missing values are introduced in Chapter 5.

2.3.3 Some Machine Learning Approaches

Using machine learning techniques, some approaches have been proposed for handling missing data in classification problems avoiding explicit imputations. In these approaches missing values are incorporated to the classifier. Treating missing values as a separate value, decision trees like C4.5 and CART were among the first algorithms to incorporate the handling of missing data into the algorithm itself (Quinlan, 1993; Breiman et al., 1984). Neural network ensemble models and also fuzzy approaches are other examples. For more detailed information on these algorithms we refer the reader to (García-Laencina et al., 2010; Breiman et al., 1984; Quinlan, 1993).

CHAPTER 3 SELECTIVE SAMPLING DESIGNS TO IMPROVE THE PERFORMANCE OF CLASSIFICATION METHODS

This chapter adresses the first research question, namely can an entropy-based selective sampling design improve the performance of classification models? This question is at the cornerstone of the thesis.

3.1 Chapter Overview

When the training of a classifier has a fixed number of observations and missing values are unavoidable, we can decide to allocate the observations differently among the variables during the data gathering phase. We refer to this situation as Selective Sampling.

One important example is Computerized Adaptive Testing (CAT). Student test data are used for training skill mastery models. In such models, test items (questions) represent variables that are used to estimate one or more latent factors (skills). For a number of practical reasons, the pool of test items often needs to be quite large, such as a few hundreds and even thousands of items. However, for model training, it is impractical to administer a test of hundreds of questions to examinees in order to gather the necessary data. We are thus forced to administer a subset of these test items to each examinee, leaving unanswered items as missing values. Hence, adaptive testing is a typical context where we have the opportunity to decide which variable will have a higher rate of missing values, and the question is whether we can allocate the missing values in a way that will maximize the model's predictive performance?

Although CAT is a typical application domain where we can apply Selective Sampling, any domain which offers a large number of features from which to train a model for classification or regression purpose is a good candidate for Selective Sampling. The datasets used in this experiment represent examples of such domains (see table 3.1 for a full list).

In this chapter[1], we investigate the effect of an informed selective sampling heuristic to improve the prediction performance of three different classifiers. The rest of the chapter is organized as follows. Below we first introduce the binary entropy function and then in section 3.3, give a brief description of the classification models used in this study. In section 3.4 our experimental methodology is explained. In section 3.5 the results are presented and finally,

[1]This study has been published at the following venue: Ghorbani, S. and Desmarais, M.C. (2013) Selective Sampling Designs to Improve the Performance of Classification Methods. In Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA), volume 2, pages 178-181. IEEE. Miami, Florida.

in section 3.6, we discuss the results and propose further studies.

3.2 Entropy

The informed selective sampling method we propose here in this study, relies on the entropy of a feature, where the probability of an attribute is estimated by the relative frequencies of its values.

The binary entropy function, denoted $H_2(x)$, is defined as the entropy of a Bernoulli process with probability of success $P(x = 1) = p$.

If $P(x = 1) = p$ then $P(x = 0) = 1 - p$, the entropy of x is given by:

$$H_2(x) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p} \tag{3.1}$$

The logarithms in the aforementioned formula are usually taken to the base 2 (See figure 3.1) (MacKay, 2003).

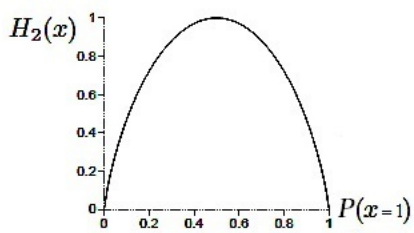

Figure 3.1 The Binary Entropy Function (MacKay, 2003)

3.3 Models

We test the hypothesis that Selective Sampling with an entropy-driven heuristic affects model predictive performance over three types of well known classifiers: Naive Bayes, Logistic Regression, and Tree Augmented Naive Bayes (TAN). They are briefly described below.

3.3.1 Naive Bayes

A Naive Bayes classifier is a simple but important probabilistic classifier based on applying Bayes' Theorem with strong (naive) independence assumptions which assume all the input

attributes are independent given its class:

$$P(c_j \mid x_1, x_2, ..., x_d) = \frac{P(c_j)}{P(x_1, x_2, ..., x_d)} \prod_{i=1}^{d} P(x_i \mid c_j) \tag{3.2}$$

Where:

$P(c_j \mid x_1, x_2, ..., x_d)$ is the posterior probability of class membership, i.e., the probability that X belongs to c_j

$P(x_1, x_2, ..., x_d)$ is the prior probability of predictors which is also called the evidence and

$P(c_j)$ is the prior probability of class level c_j

Using Bayes' rule above, the classifier labels a new case X with a class level c_j that achieves the highest posterior probability. Despite the model's simplicity and the fact that the independence assumption is often inaccurate, the naive Bayes classifier is surprisingly useful in practice.

3.3.2 Logistic Regression

Logistic regression is one of the most commonly-used probabilistic classification models that can be used when the target variable is a categorical variable with two categories (i.e. a dichotomy) or is a continuous variable that has values in the range 0.0 to 1.0 representing probability values or proportions. The logistic regression equation can be written as:

$$P = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n)}} \tag{3.3}$$

Logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target.

3.3.3 Tree Augmented Naive Bayes (TAN)

Naive Bayes classifier has a simple structure as shown in figure 3.2(a), in which each attribute has a single parent, the class to predict. The assumption underlying Naive Bayes is that attributes are independent of each other, given the class. This is an unrealistic assumption for many applications. There have been many attempts to improve the classification accuracy and probability estimation of Naive Bayes by relaxing the independence assumption while at the same time retaining much of its simplicity and efficiency.

Tree Augmented Naive Bayes (TAN) is a semi-Naive Bayesian learning method that was proposed by Friedman et al. (1997). It relaxes the Naive Bayes attribute independence

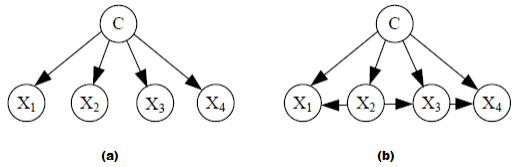

Figure 3.2 a) Naive Bayes Classifier Structure and b) TAN Classifier Structure

assumption by employing a tree structure, a structural augmentation of Naive Bayes classifier that allows the attribute nodes (leaves) to have one more parent beside the class. The structure of TAN classifier is shown in figure 3.2(b).

A maximum weighted spanning tree that maximizes the likelihood of the training data is used to perform classification. Inter-dependencies between attributes can be addressed directly by allowing an attribute to depend on other non-class attributes. Friedman et al. showed that TAN outperforms Naive Bayes, yet at the same time maintains the computational simplicity (no search involved) and robustness that are characteristic of Naive Bayes (Friedman et al., 1997).

3.4 Experimental Methodology

Our experiments have been carried out using the mentioned models and a Selective sampling design based on the entropy heuristic, the process and the datasets that are introduced below.

3.4.1 Entropy-based heuristic for Selective Sampling

We define three sampling schemes to determine missing values in order to investigate their respective effects over the predictive accuracy of the classifier models :

1. **Uniform**: Uniform random samples (Random distribution of missing values among the variables).

2. **Low Entropy**: Higher sampling rate for low entropy variables (High entropy variables will have higher rates of missing values).

3. **High Entropy**: Higher sampling rate for high entropy variables (Low entropy variables will have higher rates of missing values).

As mentioned, the entropy, or uncertainty, of a variable is derived from its initial probability of success. The probability of sampling based on entropy is a function of the $x = [0, 2.5]$ segment

of a normal (Gaussian) distribution as reported in figure 3.3. The probability of a variable being sampled will therefore vary from a scale of 0.40 at the highest, to 0.0175 at the lowest. In other words, the odds of variables being sampled is about 23 times greater at the highest level compared to the lowest level (0.40/0.0175). Variables are assigned the probability of being sampled as a function of their rank: they are first ranked according to their entropy and they are attributed a probability of being sampled following this distribution. The distributions are the same for both conditions (2) and (3), but the ranking is reversed between the two of them. For the uniform condition (1), all variables have equal probability of being sampled.

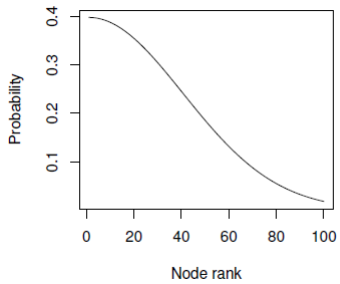

Figure 3.3 Sampling probability distribution used for the schemes 2 and 3

We have conducted a simulation study of such sampling schemes. The details of the experimental conditions and the results are described below.

3.4.2 The Process

Our simulations consist in 100-fold cross-validation runs. In each run, different training and validating sets are built based on our three schemes described in previous subsection. The proportion of total missing values inserted in the training sets is half of the data. Testing datasets contain no missing values. We compare the performance of the models on the three different sampling schemes in terms of average number of Incorrectly Classified Items (ICI) and also the average Root Mean Square Error (RMSE).

To determine whether our results are statistically significant, for each model, 2-tailed paired t-tests are run on the pairs scheme2/scheme1 and scheme3/scheme1 on the results of 100 folds.

3.4.3 Datasets

The experiments are conducted over 11 sets of real binary data. Table 3.1 reports general statistics on these datasets. The first dataset in the list, SPECT Heart, is from UCI Machine

Learning Repository (Bache and Lichman, 2013) and others are from KEEL-dataset Repository (Alcalá et al., 2010).The numbers of instances in training and validating datasets are taken to be 90 and 10 percent of the instances in the datasets respectively.

Table 3.1 Datasets - The Mean, Minimum and Maximum of the attribute entropies have been listed

	Datasets	**Attributes**	**Instances**	**Features Entropy**			**Class Entropy**
				Mean	**Min**	**Max**	
D1	SPECT Heart	22+Class	267	0.864	0.560	1.000	0.734
D2	England	100+Class	1003	0.219	0.045	0.897	0.681
D3	Ketoprostaglandin-f1-alpha	100+Class	1003	0.166	0.029	0.807	0.315
D4	Brain Chemistry	100+Class	1003	0.109	0.029	0.998	0.369
D5	Creatine-kinase	100+Class	1003	0.153	0.029	0.661	0.387
D6	Ethics	100+Class	1003	0.158	0.038	0.661	0.514
D7	Fundus-oculi	100+Class	1003	0.128	0.038	0.766	0.573
D8	Heart Valve Prosthesis	100+Class	1003	0.237	0.060	0.998	0.709
D9	Larynx	100+Class	1003	0.070	0.029	0.361	0.350
D10	Mexico	100+Class	1003	0.108	0.029	0.643	0.290
D11	Uric-Acid	100+Class	1003	0.104	0.029	0.807	0.311

3.5 Results

Figure 3.4 illustrates the way we conduct the sampling taking the Ketoprostaglandin-f1-alpha Dataset as an example. The upper-left graph reports the initial probability of each of the 100 attributes, and the other three graph report the probability of sampling the variables.

Table 3.2 reports the average percent of incorrectly classified items (ICI) for the methods based on the different sampling schemes. It also compares the performance of the methods based on average Root Mean Square Error (RMSE). As it is clear from the table, for this dataset, the performance of Naive Bayes improves under the sampling scheme 2. Logistic Regression performs better under scheme 3 and also, compared to other schemes, performance of TAN under scheme 3 is superior.

Results of conducting 2-tailed paired t-tests on the pairs scheme2/scheme1 and scheme3/scheme1 for the models on obtained results of 100 folds have been illustrated in tables 3.3 and 3.4. As the tables reflect, very small p-values show that there are very strong evidences against null hypothesis in those mentioned cases and therefore, our results, concluded from table 3.2, are statistically significant.

We have conducted similar simulations and evaluations for the other datasets. Our results show that Selective Sampling can effectively improve the performance of the classification

methods in most of the datasets. Table 3.5 reports the percent of datasets on which Selective Sampling schemes compared to uniform sampling, result in better classification performance. As it can be seen from the table, the performance of TAN in all of the datasets is improved by applying the third scheme of sampling.

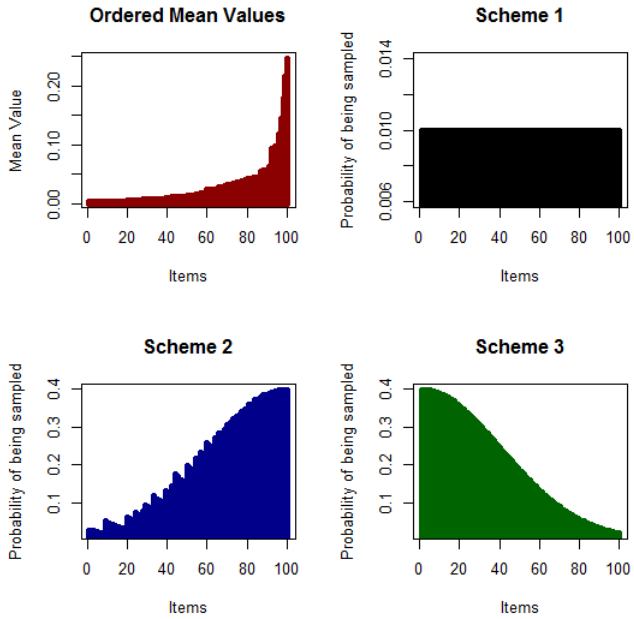

Figure 3.4 Ketoprostaglandin-f1-alpha Dataset

Table 3.2 Performance comparison for the different techniques under the different schemes of sampling for Ketoprostaglandin-f1-alpha dataset (ICI-Incorrectly Classified Items and RMSE-Root-Means-Squared-Error)

	Measure	Scheme 1	Scheme 2	Scheme 3
NB	Average Percent of ICI	4.37	3.84***	4.25
	Average RMSE	0.20	0.18***	0.19
LR	Average Percent of ICI	6.12	14.73	5.05**
	Average RMSE	0.24	0.35	0.21**
TAN	Average Percent of ICI	4.95	4.57	3.42***
	Average RMSE	0.20	0.18	0.16***

(* for 0.01<p<0.05, ** for 0.001<p<0.01 and * for p<0.001)**

Table 3.3 Results of running a paired t-test on the obtained results of 100 folds based on average Pct. ICI

	Pairs	t	Mean of the Differences	p-value
NB	Sch2/Sch1	-3.83	-0.524	0.000225
	Sch3/Sch1	-1.25	-0.117	0.213
LR	Sch2/Sch1	6.52	8.61	0
	Sch3/Sch1	-3.31	-1.07	0.0013
TAN	Sch2/Sch1	-1.49	-0.379	0.138
	Sch3/Sch1	-7.92	-1.53	0

(df=99, Confidence Interval=95%)

Table 3.4 Results of running a paired t-test on the obtained results of 100 folds based on average RMSE

	Pairs	t	Mean of the Differences	p-value
NB	Sch2/Sch1	-4.21	-0.0134	0.000056
	Sch3/Sch1	-0.993	-0.00192	0.323
LR	Sch2/Sch1	7.53	0.114	0
	Sch3/Sch1	-3.1	-0.0228	0.00253
TAN	Sch2/Sch1	-2.56	-0.0128	0.0119
	Sch3/Sch1	-8.39	-0.0356	0

(df=99, Confidence Interval=95%)

Table 3.5 Percent of datasets on which Selective Sampling classification performance results are better ($p<0.05$) than Scheme 1

	Scheme 2	**Scheme 3**	**Total**
NB	9.1%	45.5%	54.6%
LR	9.1%	45.5%	54.6%
TAN	-	100%	100%

It should be mentioned that although, for example, in 54.6% of the datasets Selective Sampling improves the classification performance of Naive Bayes, it doesn't imply that in the rest of the datasets (45.4%), Scheme 1 is the superior one; rather, as table 3.6 shows, only in 27% of the datasets uniform classification performance results are better ($p<0.05$) than just Scheme 2 and in none of the datasets it yields better results than Scheme 3.

Table 3.6 Percent of datasets on which Scheme 1 classification performance results are better ($p<0.05$) than Scheme 2, Scheme 3 or both of them

Sch.1 is better than	**Sch.2**	**Sch.3**	**Both**
NB	27%	0	0
LR	81.8%	36%	27%
TAN	63.6%	0	0

3.6 Conclusion

These results confirm that based on a heuristic that relies on attribute entropy, Selective Sampling can improve the performance of the classification methods. Selective Sampling in all of the datasets improves the performance of TAN classifier. In more than half of the datasets (54.6%) it results in better prediction performance for both NB and LR classifiers. Results also show that lower sampling rate of missing values for high entropy variables (Scheme3) brings a higher predictive performance than for the high entropy or the uniform scheme.

Further analysis and investigations are required to better explain these results. How should we explain the performance differences between the sampling schemes? What is the best sampling scheme in a given context? These are some interesting questions that are left open.

Nevertheless, this investigation shows that we can influence the predictive performance of a classifier with partial data when we have the opportunity to select the missing values. It opens interesting questions and can prove valuable in some contexts of application.

CHAPTER 4 AN ADAPTIVE SAMPLING ALGORITHM TO IMPROVE THE PERFORMANCE OF CLASSIFICATION MODELS

This chapter[1] extends the former study to assess the performance of the selective sampling heuristics without assuming the prior information, a process we refer to as Adaptive Sampling and helps guide the sampling on the fly.

4.1 Chapter Overview

Given a fixed number of observations to train a model for a classification task, a Selective Sampling design helps decide how to allocate more, or fewer observations among the variables during the data gathering phase, such that some variables will have a greater ratio of missing values than others. In previous chapter, we established that conducting an informed selective sampling which relies on features' entropy can improve the performance of some classification models. However, the results of the study were obtained given a priori information on the entropy of each variable. In realistic settings such information is generally not available. We now investigate whether the gains observed with a priori information hold when such information is not given.

The remainder of the chapter is organized as follows. We first define the term Adaptive Sampling in section 4.2 and then in section 4.3, explain our experimental methodology. In section 4.4, the results are presented and discussed. Finally, We wrap up the chapter with some concluding remarks.

4.2 Adaptive Sampling

Adaptive sampling is a technique that is enforced while a survey is being fielded—that is, the sampling design is modified in real time as data collection occurs—based on information gathered from previous sampling that has been completed. Therefore, when sampling or 'allocating' adaptively, sampling decisions are dynamically made as data is gathered.

[1]This study has been published at the following venue: Ghorbani, S. and Desmarais, M.C. (2014) An Adaptive Sampling Algorithm to Improve the Performance of Classification Models. In Proceedings of the 8th European Conference on Data Mining, pages 21-28. Lisbon, Portugal.

4.3 Methodology

Our experiments have been carried out using the same classifier models and the 11 datasets introduced earlier in Chapter 3. We have also used the three sampling schemes described in previous chapter (i.e., uniform, low entropy and high entropy sampling). The process and the details of the experimental conditions are explained below.

4.3.1 Adaptive Sampling and Seed Data

To conduct our sampling designs in an adaptive manner we start with a small seed dataset. Initial probabilities are obtained from the seed dataset and then entropy values are extracted. Then the algorithm samples feature observations based on the three different schemes. Levels of uncertainty (entropies) for all of the variables are updated based on what have been sampled so far. This process is repeated until the final sampling criterion, which in this study is to reach a fixed number of observations. Figure 4.1, shows a simple flowchart of the algorithm. In this study 3 different sizes for the seed dataset are: 2, 4 and 8 records.

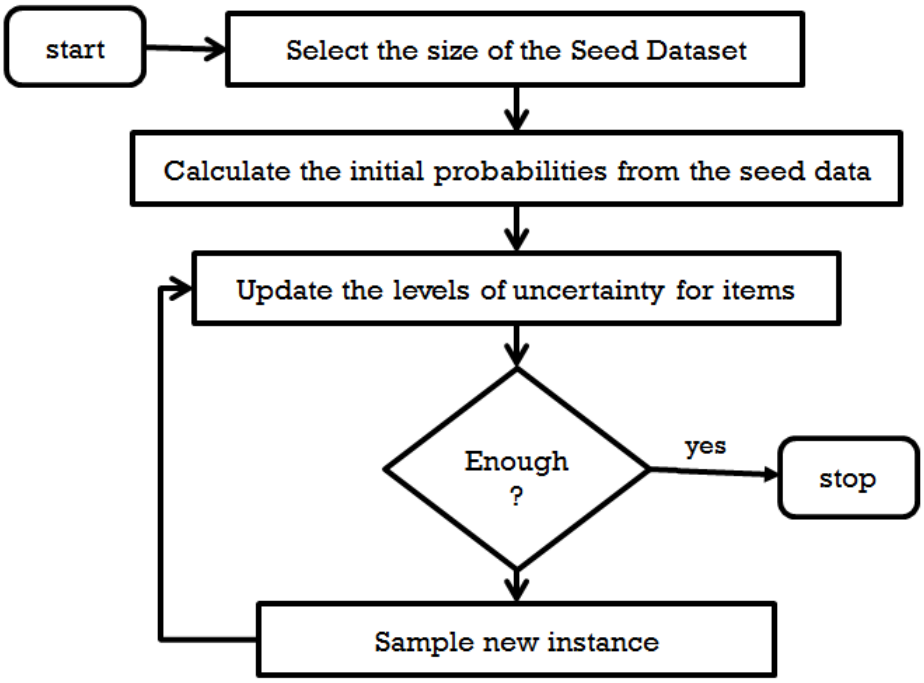

Figure 4.1 Adaptive Sampling Algorithm

4.3.2 Non-adaptive Selective Sampling

As a comparison basis for the performance of different sizes of the seed dataset we also conduct our entropy-based selective sampling schemes in a non-adaptive manner. Unlike the adaptive algorithm in which entropy values are modified in real time as data collection continues, in the non-adaptive selective sampling condition we extract the entropy values from the full dataset in hand and then conduct the three sampling schemes. This is similar to the work presented in previous chapter and provides us with another baseline for comparison.

4.3.3 Simulation Process

Our simulations consist in 100-fold cross-validation runs. In each run, different training and validating sets are built based on our three schemes described in previous subsection. The proportion of total missing values inserted in the training sets is half of the data. Testing datasets contain no missing values. We compare the performance of the models on the three different sampling schemes in terms of average number of Incorrectly Classified Items (ICI) and also the average Root Mean Square Error (RMSE).

To determine whether our results are statistically significant, for each model, 2-tailed paired Student t-tests are run on the pairs scheme2/scheme1 and scheme3/scheme1 on the results of 100 folds.

4.4 Results

Figure 4.2 illustrates the way we conduct the sampling in our non-adaptive sampling approach taking the Brain Chemistry Dataset as an example. The upper-left graph reports the entropy value of each of the 100 attributes ordered from the lowest to the highest entropy, and the other three graphs report the probability of being sampled for each corresponding attribute (item).

The results of running the adaptive algorithm with a seed dataset of size 8 over Brain Chemistry Dataset are summarized in tables 4.1 and 4.2. Table 4.1 reports the average percent of incorrectly classified items (ICI) for the methods based on the different sampling schemes. It also shows the average Root Mean Square Error (RMSE) for each of the models under the three sampling schemes. As it is clear from the table, for this dataset where the seed dataset has 8 records, the performance of Naive Bayes improves under the sampling scheme 2. Logistic Regression performs better under scheme 3 and also, compared to other schemes,

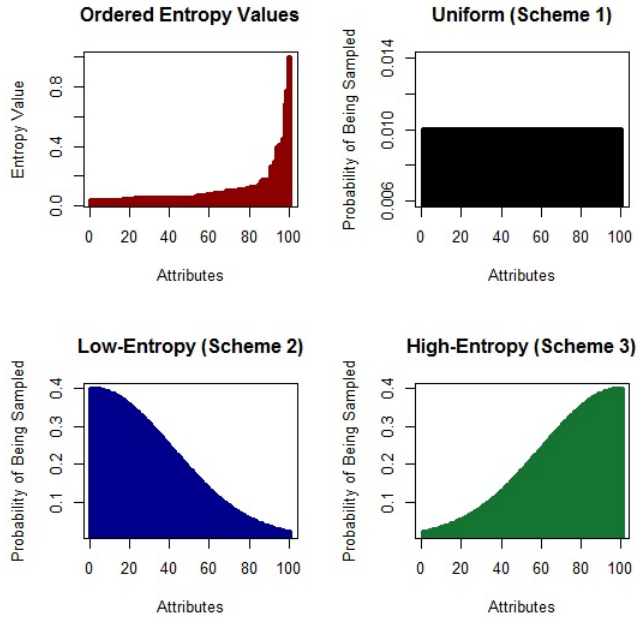

Figure 4.2 Brain Chemistry Dataset

performance of TAN under scheme 3 is superior.

Results of conducting 2-tailed paired t-tests on the pairs scheme2/scheme1 and scheme3/scheme1 for the models on obtained results of 100 folds are shown in table 4.2. As the table reflects, very small p-values show that there are very strong evidences against null hypothesis in those mentioned cases and therefore, our results, concluded from table 4.1, are statistically

Table 4.1 Performance comparison for the different techniques under the different schemes of sampling for Brain Chemistry dataset (ICI-Incorrectly Classified Items and RMSE-Root-Means-Squared-Error) where seed dataset size=8

	Measure	Scheme 1	Scheme 2	Scheme 3
NB	Average % of ICI	3.25	2.54^{***}	3.82
	Average RMSE	0.16	0.14^{***}	0.17
LR	Average % of ICI	7.58	10.08	6.53^{***}
	Average RMSE	0.27	0.30	0.25^{**}
TAN	Average % of ICI	4.31	5.24	3.27^{***}
	Average RMSE	0.18	0.19	0.16^{***}

(* for 0.01<p<0.05, ** for 0.001<p<0.01 and * for p<0.001)**

Table 4.2 RMSE difference between scheme 1 and the two other schemes for Brain Chemistry dataset. Student-t test is based on 100 random sample simulations

	Pairs	t	Mean of the Differences	p-value
NB	Sch2/Sch1	-4.3700	-0.0144	3.1e-05
	Sch3/Sch1	4.3284	0.0111	3.6e-05
LR	Sch2/Sch1	5.5064	0.0373	0.00
	Sch3/Sch1	-3.2150	-0.0177	0.002
TAN	Sch2/Sch1	1.9707	0.0075	0.052
	Sch3/Sch1	-7.1644	-0.0240	0.00

(df=99, Confidence Interval=95%)

significant.

We have conducted similar simulations and evaluations for the other datasets and the seed sizes. Tables 4.3 (a to c) summarize the results. In these tables schemes 2 and 3 are compared to the uniform sampling scheme (Sch1) based on the measure of ARMSE. The numbers in cells represent the number of datasets and those in parenthesis show the percentage of mean improvement on ARMSE value gained by applying the scheme. As it can be seen from the table 4.3-a, NB model in 45.5% of the datasets receives about 8% improvement to its prediction performance when we apply second sampling scheme with the seed size equal to 8. In general, for NB, scheme 2 is almost always better than scheme 1 in adaptive sampling approach. The table also shows compared to the third scheme, scheme one is preferable for adaptive approach.

For LR model no clear pattern emerges. But, at least it is clear from the table 4.3-b that compared to scheme 2 (which is better for only one dataset), scheme 1 brings a higher prediction performance to the classifier. We see that the sensitivity of the model to the third scheme of sampling increases when the seed size goes higher, such that we see in 27.3% of the cases, applying scheme 3 results in about 10% less ARMSE for the model compared to scheme 1 when the size of the seed dataset is 8.

For the TAN model, as table 4.3-c demonstrates, applying the third scheme of sampling in non-adaptive approach on all the datasets brings more than 13% higher prediction performance to the model. By having 8 records (less than 1% of total instances) in the seed dataset adaptive algorithm yields almost the same results as non-adaptive approach does. In none of the dataset uniform sampling is better than the third scheme of sampling, but, compared to scheme 2, uniform sampling scheme generally results in better prediction performance for TAN. Again we see a convergence in the model's performance to the case of non-adaptive approach when the size of seed dataset is 8.

Table 4.3 Number of datasets which show significant greater error (ARMSE) for each technique, under different sampling schemes, over 11 different datasets, and for different seed dataset sizes

a) Naive Bayes

	Sch1<Sch2	Sch1>Sch2	Sch1<Sch3	Sch1>Sch3
SD=2	-	4 (6.4%)	2 (5.9%)	-
SD=4	-	4 (11.7%)	5 (6.2%)	-
SD=8	-	5 (7.7%)	5 (4.8%)	-
Full	1 (10%)	4 (6.2%)	-	5(6.0%)

b) Logistic Regression

	Sch1<Sch2	Sch1>Sch2	Sch1<Sch3	Sch1>Sch3
SD=2	7 (22.6%)	1 (23.5%)	3 (12.9%)	1 (13.6%)
SD=4	6 (21.1%)	1 (16.7%)	6 (9.8%)	1 (9.1%)
SD=8	7 (21.7%)	1 (22.2%)	4 (11.2%)	3 (9.3%)
Full	9 (35.3%)	1 (16.7%)	5 (10.2%)	5 (14.3%)

c) Tree Augmented Naive Bayes

	Sch1<Sch2	Sch1>Sch2	Sch1<Sch3	Sch1>Sch3
SD=2	5 (14.2%)	-	-	10 (13.1%)
SD=4	5 (10.8%)	-	-	10 (14.3%)
SD=8	6 (10.5%)	1 (5.6%)	-	11 (12.7%)
Full	7 (14.1%)	1 (11.0%)	-	11 (13.6%)

- Schi<Schj means ARMSE(Schi)<ARMSE(Schj)
- The numbers in cells represent the number of datasets and those in parenthesis show the percentage of mean improvement on ARMSE gained by applying the scheme

4.5 Conclusion

These results confirm that Adaptive Sampling based on a heuristic that relies on attribute entropy can improve the performance of some classification methods with a 0/1 loss function. Adaptive Sampling in all but one of the datasets improves the performance of TAN classifier when we use a seed dataset of 1% or less of the total number of instances. Improvements were also obtained for the Naive Bayes classifier, but they are not systematic, and are obtained from scheme 2 instead of scheme 3. The results also show an unexpected result for one dataset, for which the uniform (scheme 1) scheme is better than scheme 2 when the entropy from the full data is taken. The Logistic regression classifier generally does better with the uniform sampling scheme, but the results are not systematic across datasets.

Further analysis and investigations are required to better explain these results. Nevertheless, this investigation shows that we can influence the predictive performance of a classifier with partial data when we have the opportunity to select the missing values. It opens interesting questions and can prove valuable in some contexts of application.

CHAPTER 5 PERFORMANCE COMPARISON OF RECENT IMPUTATION METHODS FOR CLASSIFICATION TASKS OVER BINARY DATA

We demonstrated in the previous chapters that informed sampling can bring classification performance improvement, namely for the TAN classifier. Given that informed sampling necessarily involves missing values, it is very likely to be used in conjunction with imputation methods. Imputation is a requirement for some classifiers, and it very often brings a performance improvement even to the classifier can handle missing values.

Granted that informed sampling and imputation are used in tandem, an obvious question that arises is whether the gain observed for informed sampling is added on top of the gain that we can obtain from imputation (research question 5). This question is addressed in a later chapter. First, we study in details the gains from different imputation algorithms under different sampling scheme (fixed and non fixed). This is the topic of this chapter and the next.

5.1 Chapter Overview

Three types of problems are associated with missing values in data mining (Luengo et al., 2012): (1) loss of efficiency; (2) complications in handling and analyzing the data; and (3) bias resulting from differences between missing and complete data. Furthermore, for the specific task of classification, most algorithms cannot work directly with incomplete datasets (García-Laencina et al., 2010) and therefore, reverting to imputation is often unavoidable. Missingness is particularly present for high dimensionality problems where most, if not all cases are incomplete. For example, users of a recommender system only rate a very small fraction of the available books, movies, or songs, leading to massive amounts of missing values.

Generally, imputation procedures are divided into two main categories (Farhangfar et al., 2008; García-Laencina et al., 2010; Little and Rubin, 1987):

1. **Imputation methods based on statistical analysis:** These techniques can be subdivided into two subcategories: (i) model based and (ii) quasi-randomization inference based (data driven) (Farhangfar et al., 2008). Model based methods assume that the population quantities of interests are outcomes of random attributes (variables), indexed by unknown population parameters. Quasi-randomization procedures, on the

other hand, assume that the population values are fixed, i.e., they are not governed by unknown parameters, and therefore are not the outcomes of random attributes. Statistical methods range from simple data driven methods such as mean imputation to complex model based methods that perform parameter estimation such as likelihood based algorithms.

2. **Imputation methods based on machine learning**: These methods are sophisticated procedures, which generally consist of creating a predictive model to estimate values that will substitute those missing. In contrast to statistical methods, these algorithms generate a data model from data that contain missing values, and next the model is used to perform classification that imputes the missing values. There are several options varying from imputation with K-nearest neighbors (K-nn) to imputation procedures based on auto-associative neural networks (AANN) (García-Laencina et al., 2010).

The goal of this chapter[1] is to present a comprehensive study of the impact of applying new imputation methods on classification accuracy of several leading classifiers (including TAN which has not been thoroughly studied in the literature). Furthermore, most of the investigations on imputation of missing values for classification purposes focus on a very limited scope of relatively small amounts of missing values (i.e. between 1% to 20%), with the notable exception of (Farhangfar et al., 2008). This study investigates a wider range of consistent amounts of missing data which allows for a wide range of evaluation of the quality of imputation with respect to the varying missing rates. To this end, this study provides:

- 4 different classification models including 2 generative (Naïve Bayes and Tree Augmented Naïve Bayes) and 2 discriminative models (Logistic Regression and Support Vector Machine). The models belong to major families of machine learning algorithms: NB and TAN are two probabilistic models and SVM belongs to kernel-based classifiers.

- 4 imputation methods including 2 single (missForest and Hot-deck) and 2 multiple imputation methods (based on Logistic Regression and Expectation-Maximization algorithms).

- A wide range of missing data rates per record for all the datasets (5%, 10%, 20%, 30%, 40%, and 50%).

[1]This chapter is under review at the following venue: Ghorbani, S. and Desmarais, M.C. (Under Review) Performance Comparison of Recent Imputation Methods for Classification Tasks over Binary Data. Journal of Applied Artificial Intelligence

- A number of binary test datasets: 14 datasets with large number of features.

In the context of this study, we focus on fixed number of missing values per record. One important example of such incomplete records is Computerized Adaptive Testing (CAT). Student test data are used for training skill mastery models. In such models, test items (questions) represent variables that are used to estimate one or more latent factors (skills). For a number of practical reasons, the pool of test items often needs to be quite large, such as a few hundreds and even thousands of items. However, for model training, it is impractical to administer a test of hundreds of questions to examinees in order to gather the necessary data. We are thus forced to administer a subset of these test items to each examinee, leaving unanswered items as missing values. Another example can be found in the context of medical records where we also have a limited number of medical tests available per patient.

The rest of the chapter is organized as follows. After reviewing the relevant previous work, section 5.3 briefly explains the missing value mechanism assumed in this study and presents the basis of the imputation methods and gives a brief review of the classification models used in this study. In section 5.4, the experimental framework including the methodology and the benchmark datasets is introduced. In section 5.5, the results obtained are analysed. Finally, we make some concluding remarks.

5.2 Related work

Imputation methods have been widely studied. We can track the first formal studies to several decades ago. The work of Little and Rubin (1987) laid the foundation of further works in this topic, specially in statistics. The literature on imputation methods in data mining employs well-known machine learning methods as well for their studies, in which the authors show the convenience of imputing the missing values for the mentioned algorithms, particularly for classification. The vast majority of missing values studies in classification usually analyse and compare one imputation method against a few others under controlled amounts of missing values and induce them artificially with known mechanisms and probability distributions. Some recent studies that investigated the impact of imputation on the accuracy of the subsequently performed classification are briefly overviewed below.

Farhangfar et al. (2007) develop a unified framework supporting a host of imputation methods. Their framework integrates a number of imputations methods (Naive Bayes and Hotdeck) and compares this with other basic methods (mean, Linear Discriminant Analysis, etc). Farhangfar et al. (2008) extend this study using discrete data, comparing with more classical imputation methods including three single imputation methods (Mean, Hot deck and Naive

Bayes) and one multiple imputation method (a polytomous regression based method). This study investigates the performance of the classifiers RIPPER, C4.5, K-Nearest-Neighbor, Support Vector Machine with polynomial and RBF kernels, and Naive Bayes on imputed data. The missing values are produced artificially in a wide-ranging amount for each of the datasets. According to the study the impact of the imputation varies among different classifiers and imputation is advantageous for most rates of missing values above 5%. The experimental study also shows that there is no universally best imputation method among the ones studied.

Song et al. (2008) study the relationship between the C4.5 performance (as a common machine learning technique that can tolerate missing values) and the use of the K-NN imputation method over 6 datasets of software projects. They simulate three missingness mechanisms, three missing data patterns, and five missing data proportions. The results of the study show that the k-NN imputation can improve the prediction accuracy of C4.5. This finding also agrees with Batista and Monard Batista and Monard (2003). According to Song et al, missing data mechanism and pattern and also its proportion affect the classifier and imputation method performance.

Twala (2009) empirically analyses 7 different techniques to deal with artificial missing values for decision trees over 21 real datasets. He concludes that listwise deletion is the worst choice, while the multiple imputation strategy performs better than the rest of the imputation methods considered in the study, although there is no decisively better procedure.

García-Laencina et al. (2010) evaluate the influence of imputing missing values on the classification accuracy obtained by a multilayer perceptron. They consider four imputation techniques: K-NN, SOM, MLP, and EM over one synthetic and two real datasets, varying the amount of missing values introduced. They conclude that in real-life scenarios a detailed study is required in order to evaluate which missing values estimation can improve the classification accuracy.

Luengo et al. (2010) study several imputation methods for RBFN classifiers, both for natural and artificial (MCAR—see section 5.3) missing values. From their results we see that the EC method has a good synergy with respect to the RBFN methods, as it provides better improvements in classification accuracy.

Gheyas and Smith (2010) propose a single and a multiple imputation method (the terms single and multiple imputation are explained in section Imputation Methods), both of them based on a generalized regression neural network (GRNN). Their proposal is compared with 25 imputation methods ranging from machine learning methods to several variants of GRNNs. Ninety-eight datasets are used with different missing value mechanisms. Using three models

of classifiers namely, MLP, logistic regression, and a GRNN-based classifier, their results show the advantages of the GRNN based proposal.

Luengo et al. (2012) study the performance of three different categories of classification models namely, Rule Induction Learning, Approximate Models and Lazy Learning categories under fourteen different single imputation methods. Their results show that the use of determined missing values imputation methods could improve the accuracy obtained for these models. In this study, the convenience of using imputation methods for preprocessing datasets with missing values is confirmed. They also suggest that the use of particular imputation methods conditioned on the groups is required.

5.3 Description of Investigated Approaches

The assumptions we make about the missingness mechanism and the pattern of missing values can affect which imputation method could be applied, if any. Therefore, in this section we first take a brief look at the nature of missing data as the imputation methods generally rely on specific missingness mechanism assumption.

5.3.1 The Nature of Missing Data

As mentioned, when addressing missing data it is critical to know the mechanism (cause) of that missing data and any pattern to the missing data. Three general types of missing data based on the mechanism can be derived from the literature. Each type is characterized by a level of "randomness" and "missingness", and the identification of the type of missing data ultimately leads to our ability to manage (or not manage) the missing values.

- **Missing Completely at Random (MCAR)** is the highest level of randomness. There is no dependency between missing attributes at all. In other words, data are MCAR when the probability of missing data on a variable X is unrelated to other measured variables and to the values of X itself. For this type of missingness any piece of data is just as likely to be missing as any other piece of data.

- **Missing at Random (MAR).** With MAR, the probability of missing data on any attribute does not depend on its own value, but rather relies on the values of other attributes (Liu and Lei, 2006). The propensity for missing data is correlated with other study-related variables in an analysis. It could in fact be entirely determined by other variables. As an example, suppose that a school district administers a math aptitude exam, and students that score above a certain cut-off participate in an advanced math

course. The math course grades are MAR because missingness is completely determined by scores on the aptitude test (students that score below the cut-off do not have a grade for the advanced math course) (Baraldi and Enders, 2010).

- **Not Missing at Random (NMAR)**. In this case, missing data depends on the values that are missing (Liu and Lei, 2006). In other words, the NMAR mechanism describes data that are missing based on the would-be values of the missing scores. For example, consider a self-report alcohol assessment administered to high school students. NMAR data would result if heavy drinkers are more likely to skip questions out of fear of getting in trouble (Baraldi and Enders, 2010).

Of the three missing data mechanisms, it is only possible to empirically test the MCAR mechanism (methodologists have proposed a number of MCAR tests) (Baraldi and Enders, 2010). In contrast, the MAR and NMAR mechanisms are impossible to verify because they depend on the unobserved data. That is, demonstrating a relationship (or lack thereof) between the probability of missingness and the would-be values of the incomplete variable requires knowledge of the missing values. Since, in case of the MCAR , the assumption is that the distributions of missing and complete data are the same, Farhangfar et al. (2008) and Matsubara et al. (2008) state it is only in the MCAR mechanism case where the analysis of the remaining data could give a valid inference (classification in our case) due to the assumption of equal distributions. Both of the other mechanisms could potentially lead to information loss that would lead to the generation of a biased/incorrect classifier (i.e., a classifier based on a different distribution). So, MCAR mechanism is assumed in this study.

5.3.2 Imputation Methods

This study uses two single and two multiple imputation algorithms. In addition to representing the single/multiple imputation types, they also represent three imputation mainstreams (for more detailed information on classification of imputation methods refer to (Farhangfar et al., 2008; García-Laencina et al., 2010; Little and Rubin, 1987)). These algorithms are listed in table 5.1.

Table 5.1 Imputation Methods Used in This Study

Method	Imputation Mainstream	Type
HD	*Statistical data driven (quasi-randomization)*	Single Imputation
MF	*Machine Learning based*	Single Imputation
MILR	*Model based*	Multiple Imputation
MIEM	*Model based*	Multiple Imputation

Single Imputation Algorithms

In single imputation methods, a missing value is imputed a single value. The single imputation methods used in this study are:

MissForest (MF) The recently proposed missForest method is a nonparametric imputation method for basically any kind of data. It can cope with mixed-type of variables, nonlinear relations, complex interactions and high dimensionality (Stekhoven and Buhlmann, 2012). The algorithm is based on Random Forest (Breiman, 2001). For each variable, missForest fits a random forest on the observed part and then predicts the missing part. The algorithm continues to repeat these two steps until a stopping criterion is met or a user specified maximum of iterations is reached. MissForest runs iteratively, continuously updating the imputed matrix variable-wise, and assesses its performance between iterations. This assessment is done by considering the difference(s) between the previous imputation result and the new imputation result. The algorithm stops as soon as this difference (in case of one type of variable) or differences (in case of mixed-type of variables) increase. For further details see (Stekhoven and Buhlmann, 2012).

Hot deck (HD) Generally, hot deck imputation involves replacing missing values of one or more variables of a non-respondent record (called the recipient) with observed values from a respondent record (the donor) that is similar to the non-respondent record with respect to characteristics observed by both cases. The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed. By contrast, Cold-deck imputation, selects donors from another dataset. Several different versions of hot-deck imputation method are in use in the literature. Sequential hot-deck imputes the missing values in each variable by replicating the most recently observed value in that variable. This is by far one of the fastest imputation methods commonly used for item nonresponse in survey research. Only one pass of the data is needed (Sande, 1983; Templ et al., 2011).

Multiple Imputation Algorithms

Instead of filling in a single value for each missing value, multiple imputation, as its name implies, replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Multiple imputation creates m complete copies of the dataset, each containing different imputed values (i.e. across these completed datasets,

the observed values are the same, but the missing values are filled in with different imputations). Each of these m versions is then processed identically using standard complete-data methods, and the results from the m complete analyses are combined to produce inferential results. Figure 5.1 illustrates the three distinct steps to performing a multiple imputation which are: Imputation phase, Analysis phase and Pooling phase. Based on Rubin's study (Rubin, 1987), unless the rate of missingness is exceptionally high, in most situations there is simply little advantage to producing and analyzing more than a few (3-10) imputed datasets. In accordance with this study's conclusion, we have used $m = 5$ for the range of missing rates used in the current study.

Various multiple imputation methods have been proposed by different researchers. Li (1988) and Rubin and Schafer (1990) used probabilistic Bayesian models that compute imputations from the posterior probabilities of the missing data based on the complete data. The Rubin–Schafer method assumes multivariate normal distribution of the data. On the other hand, Alzola and Harrell (1999) introduced a function that imputes each incomplete attribute by cubic spline regression given all the other attributes, without assuming that the data is modeled by the multivariate distribution. A multiple imputation environment, called multivariate imputation by chained equations (MICE), was developed by Van Buuren and Oudshoorn (1999). It provides a full spectrum of conditional distributions and related regression based methods. A more recent program was developed by Honaker, King, and Blackwell, which, Amelia II. It creates multiple imputations based on the multivariate normal model (Honaker et al., 2011). The two multiple imputation approaches used in this study are further introduced below.

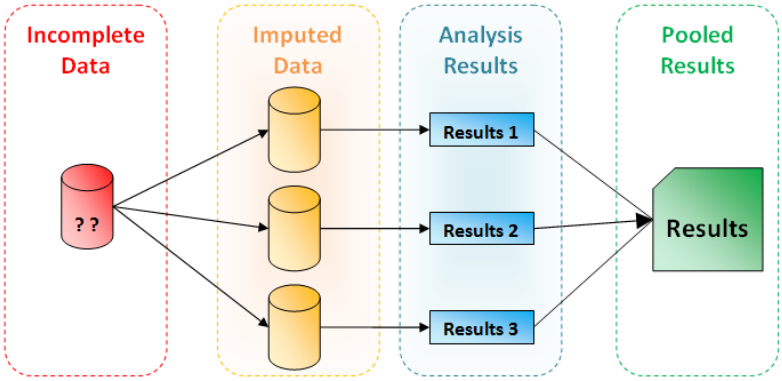

Figure 5.1 Main Steps in Multiple Imputation (here, m is assumed to be 3).

Multiple Imputation Based on Logistic Regression (MILR) MICE incorporates logistic regression, polytomous regression and linear regression, and uses a Gibbs sampler (Casella and George, 1992) to generate multiple imputation. It is furnished with a comprehensive state-of-the-art missing data imputation software package. For categorical attributes, MICE provides logistic regression (for binary attributes), polytomous logistic regression, LDA and also a simple random imputation. Our experiment uses MICE's multiple imputation based on logistic regression.

Multiple Imputation Based on Expectation Maximization (MIEM) The multivariate normality mentioned earlier in this chapter, is also assumed by Amelia II. It first bootstraps a sample dataset with the same dimensions as the original data, estimates the sufficient statistics by Expectation-Maximization (EM), and then imputes the missing values of sample. It repeats this process m times to produce the m complete datasets, where, as mentioned, the observed values are fixed and the unobserved values are drawn from their posterior distributions. In this study we use Amelia II's multiple imputation based on the EM algorithm.

5.3.3 Classifier Models

This study evaluates how the choice of different imputation methods affects the performance of classifiers that are subsequently used with the imputed data. Table 5.2 lists the 4 standard classifiers investigated in this study. We have used Weka's (Hornik et al., 2009; Witten and Frank, 2005) implementation for all the classifiers in this study. As mentioned earlier, since most of the classifiers cannot work directly with data containing missing values they need to have some internal mechanisms to handle the problem. In Weka's implementation of the classifiers an internal filter is used which globally replaces the missing values with means/modes (for nominal and numeric attributes respectively).

Table 5.2 Classifiers Considered in the Study.

Classifiers	Abbreviations
Naive Bayes	NB
Tree Augmented Naive Bayes (Friedman et al., 1997)	TAN
Logistic Regression	LR
Support Vector Machine (with RBF kernel) (Vapnik, 1996)	SVM

5.4 Experimental Methodology

Our experiments have been carried out using the mentioned classifier models and the imputation methods based on the procedure explained below. First, we describe the datasets used in the study.

5.4.1 Datasets

The experiments have been conducted over 14 sets of real binary data. Table 5.3 reports general statistics on these datasets. The first dataset in the list, SPECT Heart, is from UCI Machine Learning Repository (Bache and Lichman, 2013) and datasets D2 to D11 are from KEEL-dataset Repository (Alcalá et al., 2010). D12 to D14 come from education domain. Except for the Poly 2005, they are publicly available to their respective references. This type of data correspond to students' test results in which variables represent test items and records represent students. Since there is no class variable in such datasets, one of the variables in each dataset is randomly chosen to be the class variable used persistently across different runs as described in section 5.4.2. For all the datasets, missing values are assigned randomly into all attributes of each record (using the MCAR mechanism) in the following six amounts: 5%, 10%, 20%, 30%, 40%, and 50%.

Table 5.3 Datasets at a Glance.

Datasets		Attributes	Instances	Class Entropy
D1	SPECT Heart	22+Class	267	0.73
D2	England	100+Class	1003	0.68
D3	Ketoprostaglandin-f1-alpha	100+Class	1003	0.31
D4	Brain Chemistry	100+Class	1003	0.37
D5	Creatine-kinase	100+Class	1003	0.39
D6	Ethics	100+Class	1003	0.51
D7	Fundus-oculi	100+Class	1003	0.57
D8	Heart Valve Prosthesis	100+Class	1003	0.71
D9	Larynx	100+Class	1003	0.35
D10	Mexico	100+Class	1003	0.29
D11	Uric-Acid	100+Class	1003	0.31
D12	Fraction Subtraction Data (Robitzsch et al., 2012)	20+Class	536	1.00
D13	Poly 2005	60+Class	246	0.95
D14	Fraction algebra (Vomlel, 2002)	20+Class	149	0.72

5.4.2 Experimental Setup

Our experiments have been carried out using the procedure illustrated in figure 5.2. Firstly, for each original dataset (*Full version*) a predefined ratio of missing values is assigned to a

copy of the dataset forming *Missing-value contained version* which is subsequently imputed using the mentioned four single and multiple imputation methods resulting in four *Imputed versions* of the dataset. Finally, in a 10 fold cross validation process, the imputed datasets are used with the four classifiers: NB, LR ,TAN and SVM with RBF kernel. The classification accuracy of the classifiers are evaluated by applying the corresponding classification model on the test set, as shown in figure 5.2. The results of the above experiments are compared with two baselines in which the imputation is not performed: where (1) the classifiers are trained on data with missing values, and (2) the classification is performed on the complete data. The former experiment establishes a lower limit on accuracy, which should be improved by imputation. It should be mentioned that, in each fold, the same test dataset is used for all of the models over all the imputed and non-imputed datasets.

Therefore, using 14 datasets, six different amounts of missing values, four classifier models and four imputation methods, the mentioned procedure gives us a total of $14 \times 6 \times 4 \times 4 = 1344$ experiments in each run of a 10 fold cross validation process. Additionally, $14 \times 6 \times 4 = 336$ and $14 \times 4 = 56$ experiments are performed with the missing-value contained and complete datasets respectively in each fold.

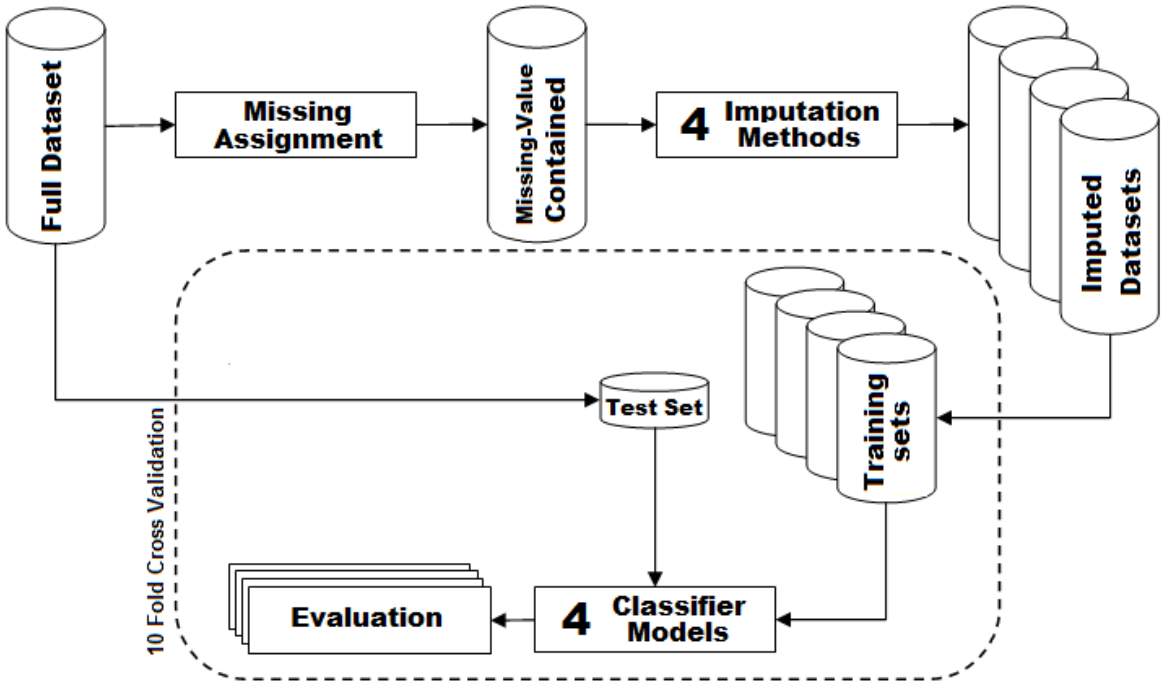

Figure 5.2 The General Procedure of the Experiments.

5.5 Results

In this study the performance of the classifiers has been assessed using a zero–one loss function, which is commonly used to evaluate the classification performance in machine learning. Although some datasets may assume different costs for their classification decisions, we assume a uniform cost for all the classes in order to be able to compare the results across different datasets.

Global Classification Accuracy as a Function of Missing Rate

Figure 5.3 gives a comparison between different imputation methods, averaged over all classification models and datasets. It illustrates the percent reduction in classification residual error based on average classification accuracy compared with the baseline. The improvements are reported for each imputation method against the different ratios of missing values. The base line (level 0) belongs to the classification on data that contains missing values. As can be seen from figure 5.3, imputation improves the performance of the classification task in general. Although, this is not the case for HD method when the missing rate is less than 20%. The largest and most consistent improvements are obtained from the MF and MIEM methods: their improvement substantially increases as the rate of missing values increases.

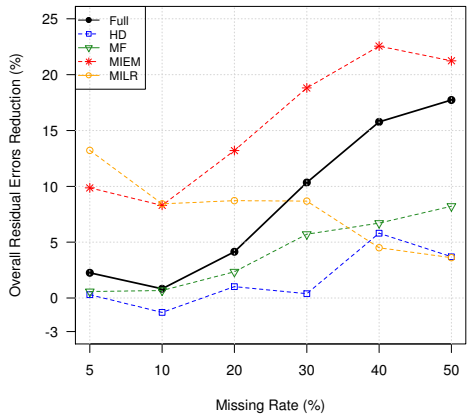

Figure 5.3 Comparison between Different Imputation Methods (Averaged over the 4 classifiers on all the datasets).

Classifier Performance under the Different Imputation Methods

Figures 5.4 and 5.5 and table 5.4 show a more detailed analysis of the results with respect to the different classifiers. They report the performance of each individual classifier under

the different imputation methods as a function of the rate of missing values. Figure 5.4 shows the classification improvement with imputed data in terms of reduction in residual errors compared to the condition of directly applying the data with missing values for LR, NB, TAN and SVM classifiers. Figure 5.5 contains notched box plots of each combination of imputation method and model, for each of the 6 levels of imputation. It provides a more detailed understanding of the performance improvements and allows a general view of the statistical significance of these improvements. Individual points in these plots represent improvement for a single run and a single dataset, and each bar contains quartiles and the median, and individual data points are outliers. It should be noted that the level 0 in the plots represents the classification using the missing-value contained datasets, as a baseline for the comparisons. Table 5.4 reports the performance of each imputation method from another perspective. It lists the percentage of the datasets on which applying a given imputation method improves the classification accuracy for each model at different missing rates. Insightful information can be learned from the table, for example when it comes to deciding about choosing an imputation method the table can tell us which of the methods is more likely to be useful with a given classifier at a given amount of missing values.

The results show that the impact of the imputation varies for different classifiers. As can be seen from the LR graph in figure 5.4, MIEM results in a substantial reduction of residual errors that increases with higher rates of missing values. Classification with imputed data at 40-50% missing rate using MIEM results in more than 36% reduction in residual errors compared to the classification on the data with missing values as the base line. Figure 5.5 shows that in case of applying the MIEM imputation method the improvement in classification accuracy for LR is statistically significant. Figures 5.4 and 5.5 also show that MILR imputation provides significant improvements for LR too. Although, the performance of MILR for the higher volumes of missing values decreases (in fact, this is the case for all the considered classifiers as the graphs show), it brings the highest reduction in residual errors to LR for smaller amounts of missing rate. For missing rates above 30%, the graphs also suggest that HD imputation method can provide significant improvements when LR is subsequently used, while using the other single imputation method, MF, results in statistically insignificant differences. Another important point to consider, as can be observed from the table 5.4, is that there is the same pattern when it comes to count the number of cases (the datasets) on which applying a given imputation method improves the classification accuracy of LR so that for low amount of missing rates (5 to 20 percent) MILR and for the higher rates (30 to 50 percent) MIEM plays the role of the most reliable imputation method to be used with LR. (e.g. MILR at 85.7% of the datasets when missing rate is 30% and MIEM at 92.9% of cases when it is higher than 30% improves the accuracy of LR.) The last point with LR is that

it is highly responsive to the considered imputation methods so that our results show when the amount of missingness in the data goes higher than 30%, For each dataset, there is at least one imputation method that brings a significant improvement to the LR classification accuracy.

Table 5.4 Percent of all the datasets on which applying the method improves the classification accuracy of the classifiers.

Classifier Model	**Missing Rate (%)**	Imputation Method				**Classifier Model**	**Missing Rate (%)**	Imputation Method			
		HD	**MF**	**MIEM**	**MILR**			**HD**	**MF**	**MIEM**	**MILR**
LR	5	42.9	50.0	71.4	64.3	**TAN**	5	64.3	71.4	57.1	64.3
	10	50.0	42.9	64.3	71.4		10	64.3	71.4	50.0	57.1
	20	50.0	78.6	78.6	85.7		20	57.1	50.0	64.3	35.7
	30	57.1	42.9	85.7	64.3		30	78.6	92.9	78.6	57.1
	40	85.7	57.1	92.9	64.3		40	92.9	71.4	78.6	50.0
	50	85.7	64.3	92.9	71.4		50	85.7	92.9	71.4	50.0
NB	5	7.14	28.6	50.0	35.7	**SVM**	5	7.14	35.7	85.7	92.9
	10	21.4	42.9	28.6	50.0		10	0.0	28.6	85.7	78.6
	20	21.4	42.9	21.4	21.4		20	14.3	42.9	78.6	57.1
	30	7.14	57.1	50.0	35.7		30	0.0	57.1	78.6	50.0
	40	35.7	78.6	50.0	28.6		40	7.14	57.1	64.3	42.9
	50	35.7	64.3	21.4	28.6		50	14.3	57.1	64.3	57.1

The graphs of figure 5.4 illustrate that the average performance improvement of classification of the full datasets over the missing-value datasets increases as a function of missing rate. This is the case for all of the models except for NB for which we see almost the same performance of the model for the both complete and incomplete datasets. It seems this leaves little room for imputation methods to help provide the classifier with an improvement in classification accuracy. One explanation for this result is that NB does well with small datasets and therefore imputation does not bring any additional improvement as the peak performance is already obtained with the observed, non imputed cases. It can only decrease if imputation adds noise in the data, as seem to be the case for MILR. The results also show that MF method brings a little improvement in classification accuracy (2.9% to 4.3% reduction in residual errors) for larger amounts of missing values (when missing rate is above 30%). Another point to mention is that in case of NB classifier, at each of the considered amounts of missing rates we can find some datasets (at least 2) on which applying the imputation methods, no matter what method is chosen, does not make a positive difference to the classification accuracy. (Refer to table A.3 in Appendix A)

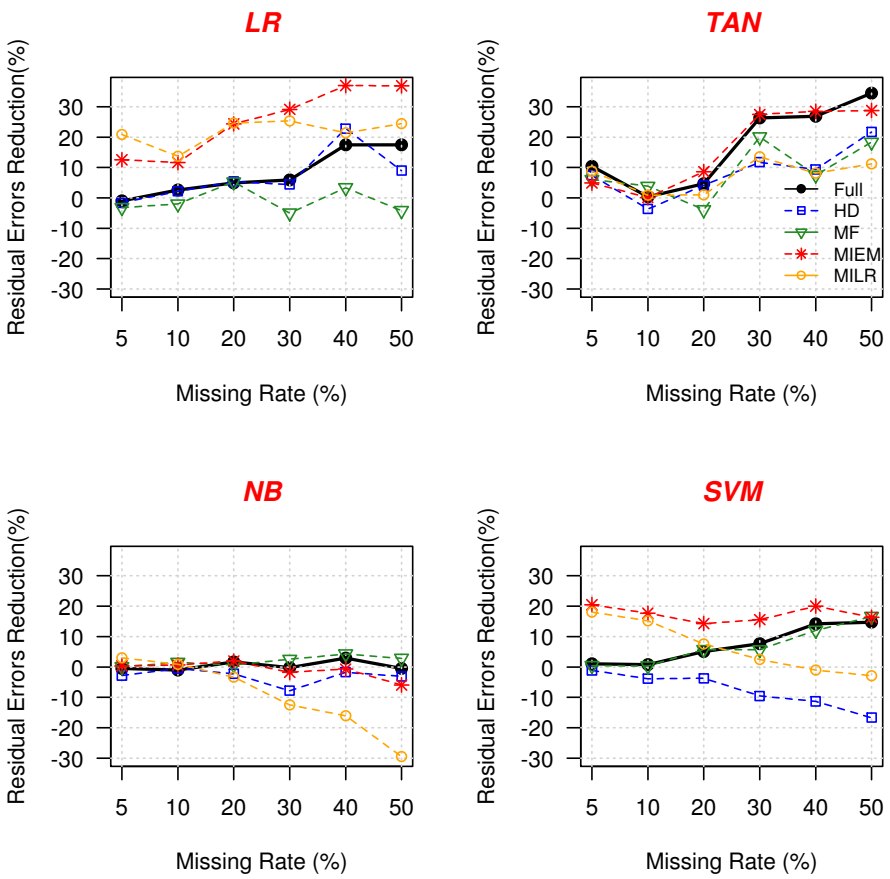

Figure 5.4 Classification Improvement for the Logistic Regression (top left), Tree Augmented Naïve Bayes (top right), Naïve Bayes (bottom left), and SVM with RBF Kernel (bottom right).

The results for TAN classifier show a significant improvement in classification accuracy on imputed data using HD, MF and MIEM methods for missing rates above 20%. For smaller rates of missing values, as can be seen from figure 5.5, imputation results are not statistically significant. Among the imputation methods while MIEM is shown to bring the highest improvement with TAN in tandem (on average, more than 27% reduction in residual errors), MF is the one that can be reliably applied on the highest percentage of the datasets (about 93% when MR is 50%) to improve the accuracy of TAN. The figure 5.5 shows that the improvements of TAN performance resulting from MILR imputation method are not statistically significant. Similar to LR, the TAN model is highly responsive to the considered imputation methods. Our results show that when the missing rate is above 20%, for each dataset we can find at least one imputation method to be used with TAN in order to make a significant improvement in the classification accuracy.

Finally, as figure 5.4 illustrates, the SVM classifier with RBF kernel benefits from imputation

too. As can also be seen from figure 5.5 and table 5.4, the large majority of the datasets (93%) show significant improvements at the 5% when MILR is applied prior to the subsequent classification. Table 5.4 confirms MIEM plays the same role at all other considered rates of missing values. Results also show that at all the considered missing rates, the highest amounts of average improvement due to imputation, are observed from applying MIEM method and we see the improvements are statistically significant at 5%, 20% and 50% of missing data (Tables A.1 and A.2 in Appendix A, provide some more detailed information). These results suggest that MIEM is the best imputation method to be used with SVM (with RBF kernel). Although, at some amounts of missing values, we can find 1 or 2 datasets on which using HD method helps improve the accuracy of the classifier, as it is clear from the figures 5.4 and 5.5, on most of the datasets it results in deterioration in prediction performance of the classifier. Results also show that MF method for remarkable number of the datasets brings a better performance for the classifier. For SVM, we noticed that no matter what the missing rate is, applying the imputation methods does not bring any improvement to the performance of the classifier on the two specific datasets: D6 and D8 (Refer to table A.3 in Appendix A).

5.6 Conclusion

Missing data is a common nuisance in many real-world applications of data mining and, in particular, in classification. Due to the high dimensionality of some datasets, we can even expect to find no complete cases at all in the dataset.

The main objective of this work is to evaluate the effect of missing data imputation on the accuracy of subsequent classification. To this end, we performed a comprehensive experimental study, which includes four single and multiple imputation methods that were used to impute missing values in 14 binary datasets. The imputed data were used to perform classification with four different classifiers to investigate the effect of imputation on the classification performance. We also considered imputations for six different amounts of missing data (i.e., 5%, 10%, 20%, 30%, 40%, and 50%) per record for each of the datasets, and compared the results obtained from classification on imputed data with the ones on missing data.

This experimental study shows that imputation with the tested methods on average improves classification accuracy when compared to classification without imputation.This agrees with the results presented by former studies (e.g., see (Farhangfar et al., 2008)). Analysis of the obtained results with respect to the classifier models used in this study shows that in case of:

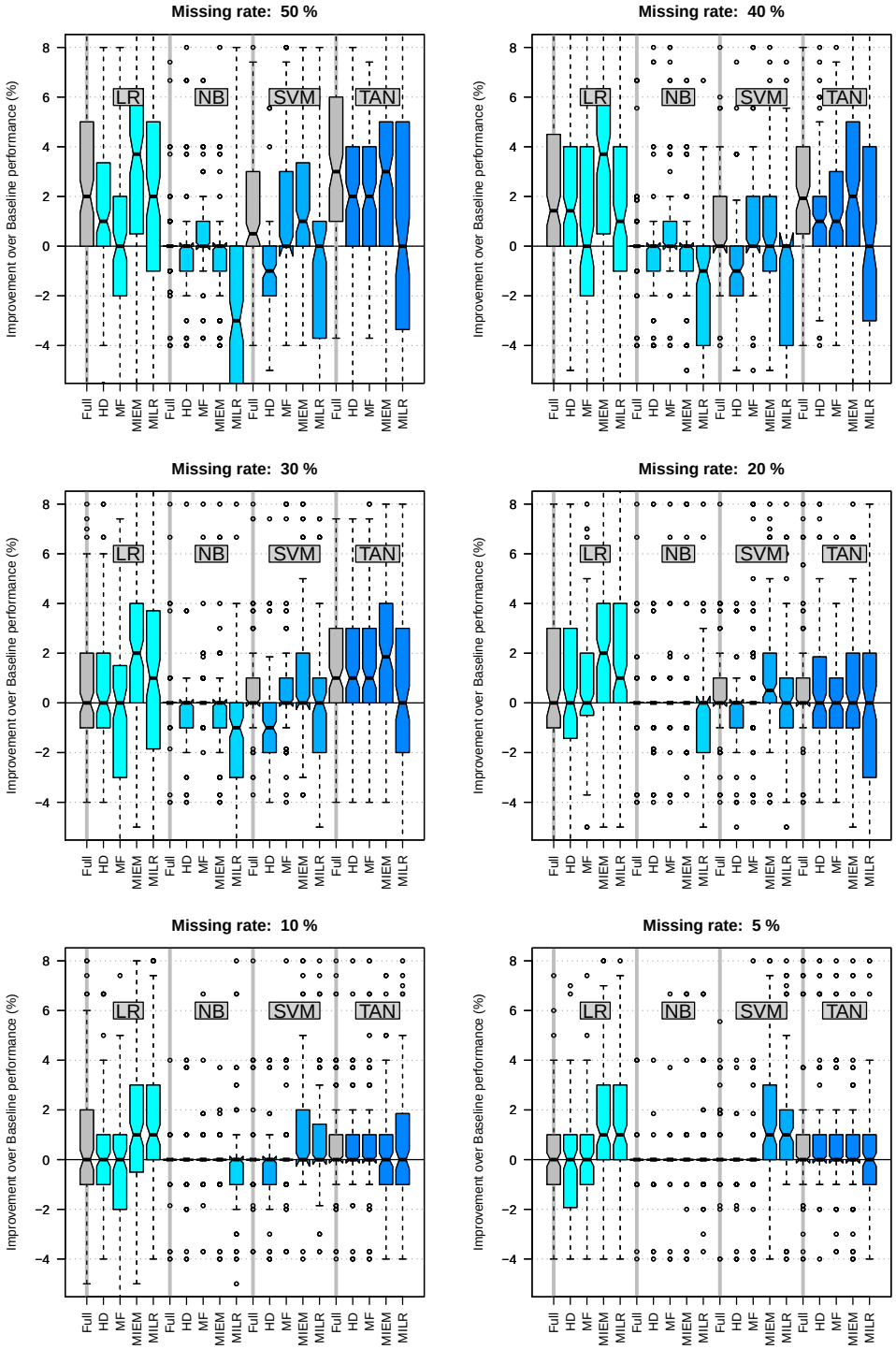

Figure 5.5 The plots represent the distribution of paired differences between the classification accuracy using each of the imputation methods and directly applying the data with missing values for each classifier.

- **LR**: the best imputation methods to be used with the classifier in tandem are MILR (when missing rate is 5% to 20%) and MIEM (when missing rate is above 20%) resulting in statistically significant improvements in classification accuracy (more than 36% reduction in residual errors). HD can be helpful too for higher amounts of missing values (above 30%).

- **NB**: the classifier shows almost the same performance on complete and incomplete data records. Although, this agrees with the conclusion made by Farhangfar et al. (Farhangfar et al., 2008) which states that NB is missing data resistant, our results show the new imputation method, MF, can bring an improved prediction performance to the classifier when the missing rate is above 30%.

- **TAN**: HD, MF and MIEM provide the classifier with statistically significant improvements in classification accuracy when missing rate is above 20%. While MIEM results in the highest amount of improvement (more than 27% reduction in residual errors), MF is the most reliable imputation method for the classifier as it improves the prediction performance of the classifier in majority of the datasets at the different considered missing rates (e.g., in 93% of the datasets when missing rate is 50%).

- **SVM** (with RBF kernel): MIEM is the best imputation method as it results in the highest amounts of improvements in accuracy of the classifier. These improvements are statistically significant at 5%, 20% and 50% missing rates. For most trivial amounts of missing data (less than 10%), MILR is the most reliable imputation method.

LR and TAN are highly susceptible to the imputation methods at higher amounts of missing values (for TAN when missing rate is above 20% and for LR when it is more than 30%). This is not the case for NB and SVM as at each considered amount of missing rates we can find a few datasets on which applying the imputation methods (no matter which one) does not improvement the classifier's accuracy. Based on results we can also conclude that:

MF seems to work best with TAN although; it can also bring an improved prediction performance to all of the other classifiers over remarkable number of datasets at different missing rates. **HD** can be helpful for TAN and LR (when MR>30%) but not for NB and SVM. The amount of improvement in classification accuracy resulting from applying **MILR** decrease as missing rate increases. Except for NB, **MIEM** is the best method to be used with all the classifiers as it results in higher amounts of improvement in accuracy of the classifier and again, except for the case of NB, these improvements are increasing against increasing amounts of missing values.

CHAPTER 6 IMPACT OF FIXED VS. VARIABLE SAMPLING RATE PER RECORD OVER THE PERFORMANCE OF IMPUTATION METHODS IN CLASSIFICATION TASKS

This chapter is the second one that evaluates the performance gains from imputation. Whereas the previous one focused on comparing a number of imputation techniques, we focus here on the impact of fixed vs. variable sampling rates on imputation (RQ4). This question arises because some of our studies have been done with a fixed sampling rate and we whish to assess whether the results can generalize to non-fixed sampling rates, which is the de facto standard sampling scheme used in assessing the impact of imputation on classifier performance.

6.1 Chapter Overview

Incomplete data is an unavoidable problem when dealing with high dimensional real world datasets and many data mining algorithms cannot work directly with incomplete datasets, and therefore reverting to imputation is often unavoidable.

The majority of missing values studies of classification tasks usually analyze and compare one imputation method against a few others used in tandem with given classifier models under predefined amounts of randomly distributed missing values. In most if not all of these studies, the number of missing values per record varies and follows a Gaussian distribution (e.g., see (Gheyas and Smith, 2010; Luengo et al., 2012; Matsubara et al., 2008; Twala, 2009; Farhangfar et al., 2008)). But in many contexts, sampling observations can be chosen, and one choice is to assign a fixed rate of observations per record. Observations are chosen at random, but the same number is observed per record. This results in records with the same rate of missing values. We will refer to such approaches to sampling as *fixed rate sampling*, and to the well studied alternative approach of sampling at random across records as the *variable rate sampling*.

Examples of fixed or variable rate sampling emerge from different domains such as Computerized Adaptive Testing (CAT) and recommender systems. In CAT, where we find applications of logistic regression models (Baker and Kim, 2004), the pool of test items can reach the thousands. Because asking a large amount of questions to a single student is impractical, we face missing values when training the models to be used for CAT. But the amount of questions per student can be either fixed or variable, depending on the experimenter's choice or on contextual factors.

Another domain where we can find fixed vs. variable sampling is found in recommender systems, where initial suggestions provide seed data to construct the user profile (Golbandi et al., 2010).

This chapter[1] investigates how the choice of the two mentioned random sampling schemes can affect the prediction performance of a given classifier model under a given imputation method at different amounts of missing values (given equal number of observations in total).

The remainder of the chapter is organized as follows. Section 6.2 gives a brief review of the imputation methods and the classification models used in this study. In section 6.3, the experimental framework, including the methodology and the benchmark datasets, is introduced. In section 6.4, the results obtained are reported and discussed. Finally, we make some concluding remarks.

6.2 Description of the Imputation Methods and the Classifiers

The effect of fixed vs. variable sampling is investigated over five imputation methods including the four methods introduced in previous chapter (i.e., HD, MF, MILR and MIEM) and Mean imputation method which imputes the missing values with the mean for numeric data or the most frequent value (mode) for nominal data of the corresponding attribute.

We focus on three standard classifier models:

- Naive Bayes (NB)
- Tree Augmented Naive Bayes (TAN) (Friedman et al., 1997)
- Logistic Regression (LR)

We have used WEKA's (Hornik et al., 2009; Witten and Frank, 2005) implementation for all the classifiers in this study.

6.3 Experimental Methodology

The five imputation methods and the three classifier models are orthogonal dimensions and together they create 15 different conditions over which we study the effect of fixed vs. variable

[1]This study is under review at the following venue: Ghorbani, S. and Desmarais, M.C. (Under Review) Impact of Fixed vs. Variable Sampling Rate per Record over the Performance of Imputation Methods in Classification Tasks. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery.

sampling. We now describe the sampling approaches used in the study and the general experimental procedure.

6.3.1 Rate and Distribution of Missing Values

Using the MCAR mechanism, missing values are assigned randomly into all attributes of all the 14 datasets described in Chapter 5, in four proportions: 10%, 30%, 50% and 70% according to 2 different sampling approaches:

- **Fixed** Rate of Observations per Record (Fixed): each record has the same amount of missing values.
- **Variable** Rate of Observations per Record (Variable): records have different amounts of missing values. The distribution of the number of missing values follows a Binomial distribution $Bin(k; n, p)$ where k is the number of missing values of the corresponding fixed rate sampling, n is the number of variables to sample per record, and p is the proportion of missing values.

6.3.2 Experimental Setup

Our experiments have been carried out using the procedure illustrated in figure 6.1. For each original dataset (Full version) as explained above, using the sampling approaches and predefined ratios, two samples, one with *Fixed* and the other with *Variable* rate of missing values (per record), are created which are subsequently imputed using the mentioned five single and multiple imputation methods resulting in five imputed versions of the dataset for each pattern of data missingness. Next, in a 100 fold cross validation process, the imputed samples are used with the 3 classifiers: NB, LR and TAN. Each classifier's accuracy is evaluated by applying the corresponding classification model on the test set, as shown in figure 6.1. The same test dataset is used for all of the models over all the imputed datasets in each fold.

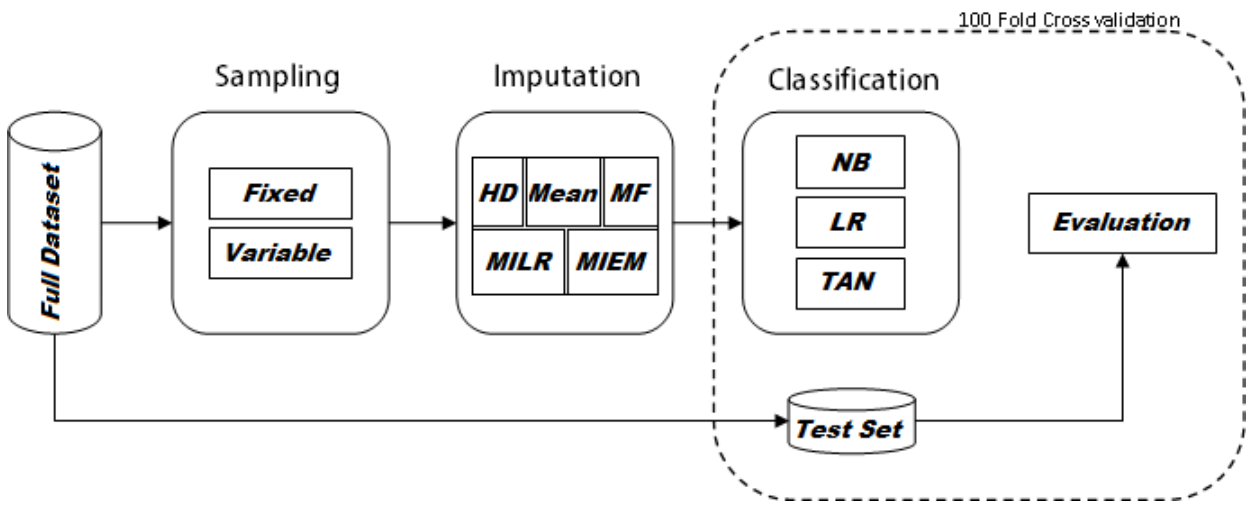

Figure 6.1 Experimental Procedure.

6.4 Results

In this study the performance of the classifiers has been assessed using a zero–one loss function. Although some datasets may assume different costs for their classification decisions, we assume a uniform cost for all the classes in order to be able to compare the results across different datasets.

Average Accuracy Differences

We first assess the effect that the fixed vs. variable sampling schemes can have on classifier performance across different imputation methods. The measure used for this purpose is the average classification accuracy of the models under different imputation methods across different proportions of missing values. Figure 6.2 shows logit of the accuracy as a function of different missing values ratios and for each combination of model and imputation with 95% confidence interval as shadows. Note that we use the logit transformation ($logit(x) = 1/(1 + e^{-x})$) because the $[0, 1]$ scale of Accuracy can create a highly skewed distribution at the extremes which does not lend itself to computing averages and confidence intervals, nor to show gains in error reduction.

As can be seen from figure 6.2, TAN is the model for which the choice of fixed vs. variable sampling schemes makes the clearest difference at the higher missing value rates and HD, Mean and MF are the most affected methods with the model.

These results prompts us to investigate further the results at the dataset level.

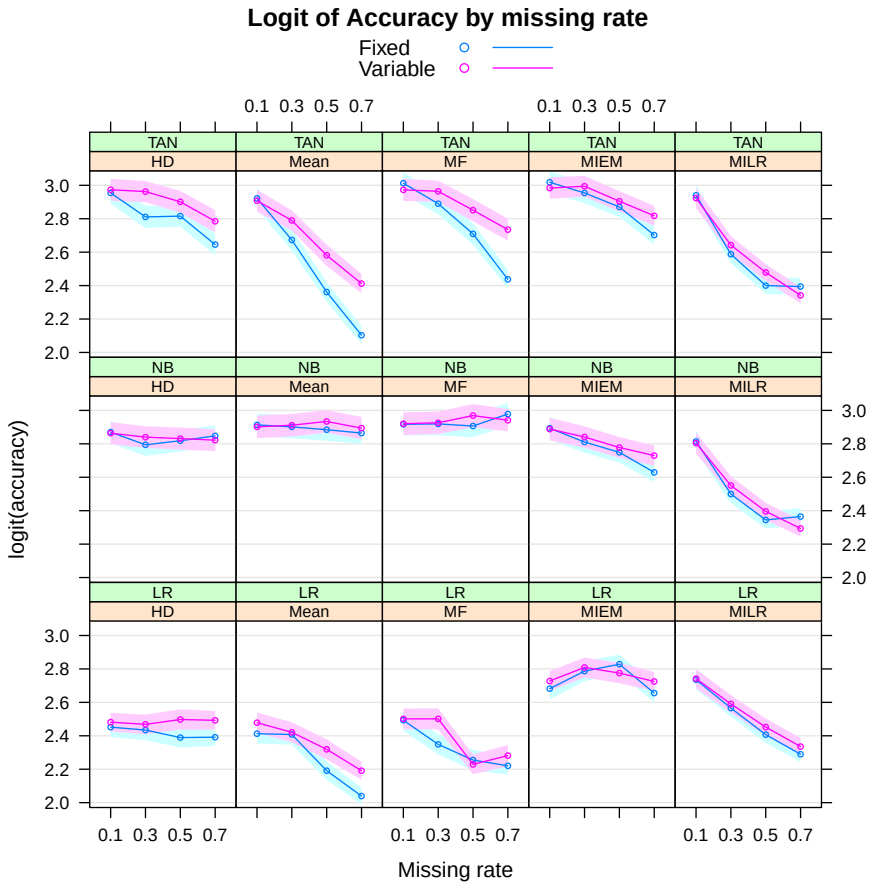

Figure 6.2 Average classification accuracy of the models under different imputation methods over fixed vs. variable sampling schemes (averaged over the 14 datasets and 100 runs).

Individual Dataset Differences

Figure 6.2 shows a trend in imputation gain generally in favor of the variable sampling scheme, particularly for the TAN model. Concurrently, the amplitude of the differences increase as a function of the rate of missing values.

To further investigate the amplitude effect and why the TAN shows a trend, we pursue the analysis over individual datasets. For each dataset, we report the outcome of a hypothesis test over the performance of fixed vs. variable sampling. The tests are done at each missing values rate over 100 cross validation runs using random train-test sampling. The results are reported in figure 6.3. The percentage of datasets for which the performance of variable (orange) or fixed (blue) sampling is significantly better is shown as a function of the missing values rate and for each classifier. Since the samples may be non-normal, we use a non-parametric equivalent of paired Student t-test in our analysis; significance level is $p < 0.05$

and based on a one-tailed Wilcoxon Signed-Rank test.

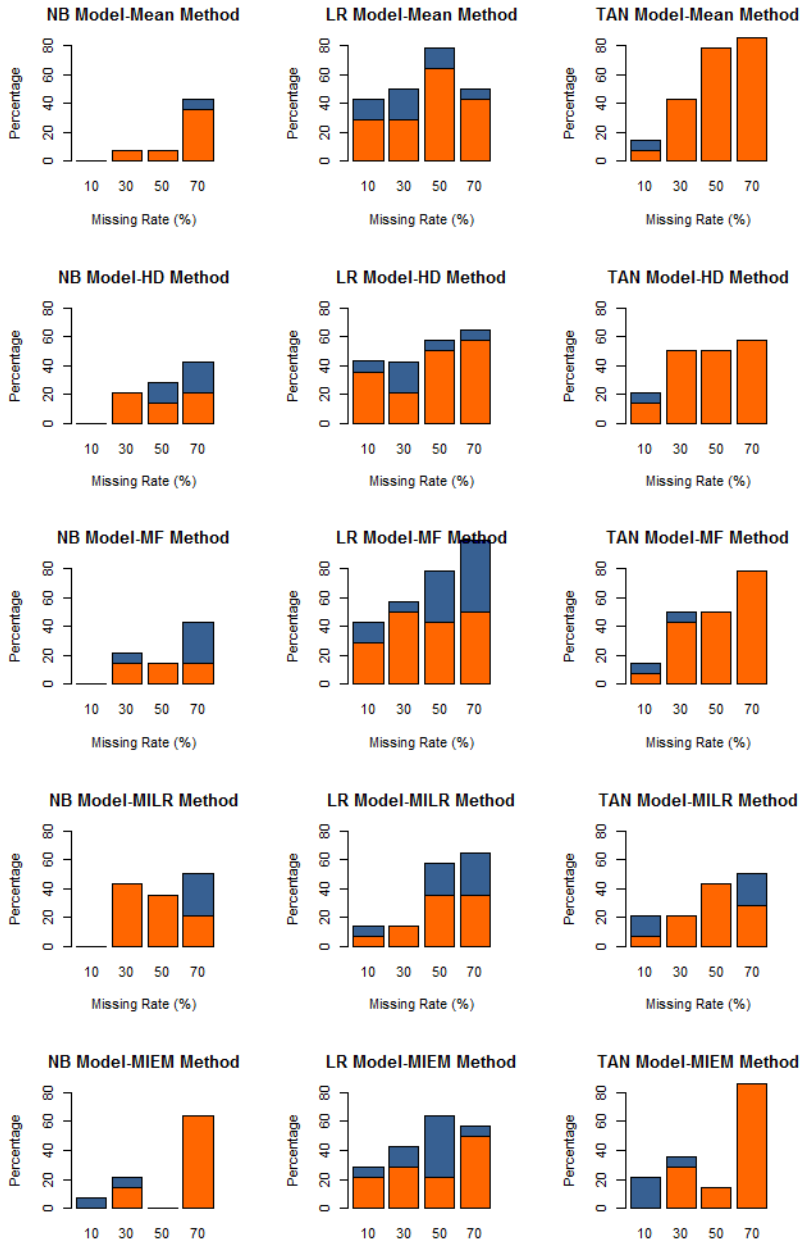

Figure 6.3 Orange) percentage of datasets on which the accuracy of a given model under a given imputation method over the variable sampling scheme is significantly higher than its accuracy under the same imputation method over the fixed scheme. Blue) similarly, shows the same percentage for the fixed sampling scheme against the variable scheme (Based on one tailed Wilcoxon Signed-Rank test at $p < 0.05$).

The results show that variable sampling outperforms fixed sampling for most of the individual datasets, as evidenced by the larger proportion of the orange color compared to the blue in figure 6.3. However, and in accordance with figure 6.2, the trend is smaller for NB and higher

for TAN. As can be seen from figure 6.3 the trend is also higher for mean method among the imputation methods studied.

Adverse Patterns of Results in Imputation Methods

We conclude the presentation of the results with a final and important perspective of the results, because one more question needs to be addressed: given that imputation of the missing values from the two sampling schemes yield different results for the classifiers and the imputation methods considered (as shown by figures 6.2 and 6.3), how does this affect the conclusions from previous studies?

As mentioned earlier, previous research has focused almost solely on the variable sampling scheme, but our analysis show that under a fixed sampling scheme, these conclusions may not hold. It is important to note that imputation of missing values of the two sampling schemes results in two possible patterns of effects over the predictive accuracy of a given classifier:

1) *Consistent (or Same Direction) Effects*- They either improve, or deteriorate the model's performance and more importantly,

2) *Inconsistent (or Opposing Direction) Effects*- Cases for which a given imputation method brings a significant improvement over variable scheme but deteriorates or does not make a difference to the model's accuracy over fixed scheme and vice versa. Here are two examples of the issue:

As can be seen from figure 6.4 –A), although MF and MILR improve the accuracy of LR model over dataset D10 for variable (orange) scheme significantly (at $p < 0.05$ based on a one-tailed Wilcoxon Signed-Rank test over 100 runs), they deteriorate the performance of the classifier for fixed (blue) sampling. Similarly, as illustrated in figure 6.4–B) while all the imputation methods bring significant improvements (at $p < 0.05$) to the predictive performance of TAN model over dataset D13 for variable sampling, they make no difference or even deteriorate the accuracy of the classifier on fixed scheme. Error bars show the confidence interval (95.0%) and the vertical axes illustrate the average gain defined as the improvement in accuracy over a baseline:

$$Gain_{i,j} = \frac{Acc_{i,j} - Acc_{base,j}}{Acc_{base,j}} \tag{6.1}$$

where $i \in \{HD, MF, MILR, MIEM\}$ and $j \in \{Fixed, Variable\}$

As noted earlier, since most of the classifiers cannot work directly with data containing missing values they need to have some internal mechanisms to handle the problem. In

Weka's implementation of the classifiers an internal filter is used which globally replaces the missing values with means/modes (for numeric and nominal attributes respectively). Therefore, the accuracy of the model under mean imputation method is considered as the baseline (denominators of the formula above).

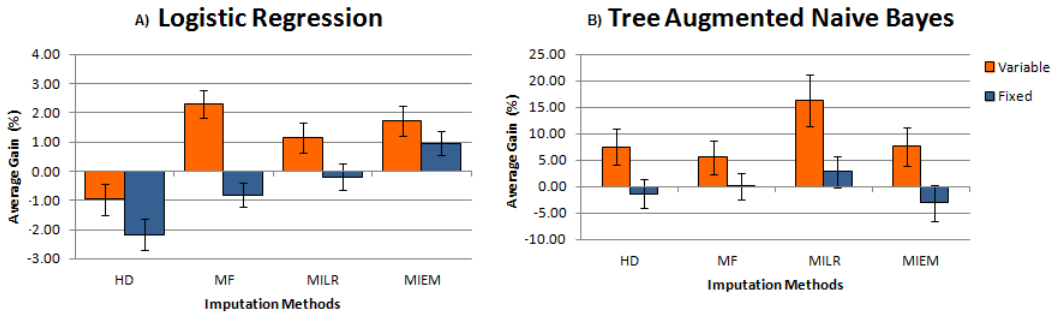

Figure 6.4 Inconsistent patterns of imputation results over fixed vs. variable sampling for: A) LR over dataset D10 at missing rate 30% and B) TAN over dataset D13 at missing rate 10%.

In order to quantify and better study the inconsistent effects of each of the imputation methods over fixed vs. variable sampling, we aggregate the pattern over datasets for each of the classifiers, as reported by figure 6.5. Again significance level is $p < 0.05$ and based on a one-tailed Wilcoxon Signed-Rank test over the results obtained from 100 simulation runs on each dataset and for each missing value rate. As the graphs in figure 6.5 show, in case of LR and TAN classification, MIEM generally yields the least number of cases with inconsistent patterns of imputation results, which is also decreasing as the missing rate increases. On the other hand, we see the largest percentages are generally resulted from MF imputation method, particularly for LR, ranging between 20% at the lowest missing data proportion and 60% at the highest.

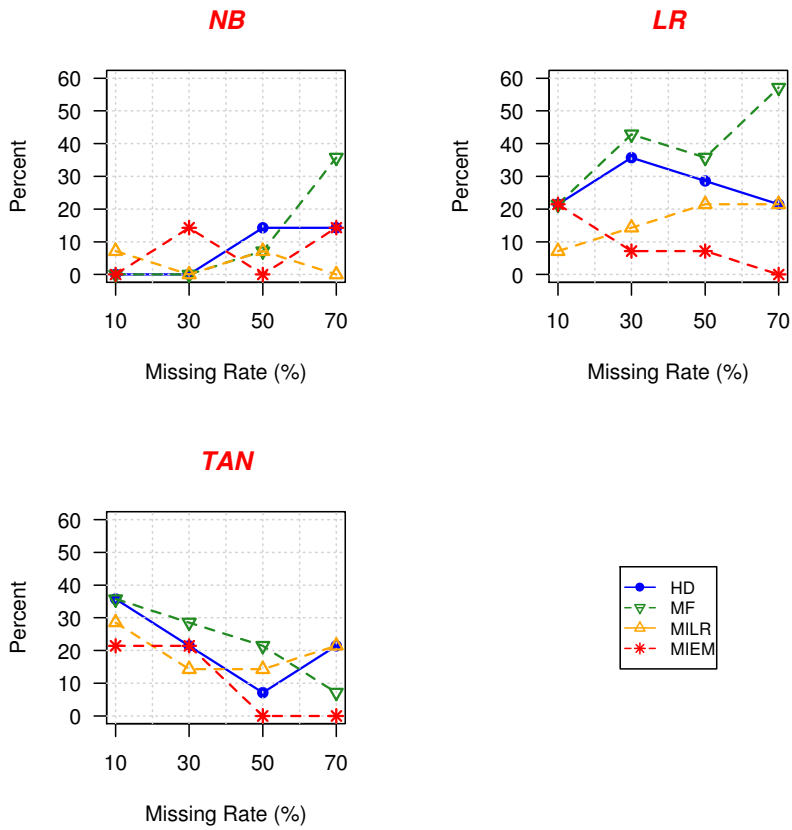

Figure 6.5 Percentage of datasets for which imputation results in inconsistent patterns of effects over variable vs. fixed sampling for each classifier.

6.5 Conclusion

The treatment of incomplete data is an important step in the pre-processing of data mining particularly in classification tasks. The main objective of this study is to investigate whether the choice of two different random sampling schemes namely, fixed and variable rates of observations per record (given equal number of observations in total), can affect the prediction performance of a given classifier model under a given imputation method.

To this end, we performed an empirical study of the classification accuracy gain obtained from different imputation methods under the two sampling schemes.

According to the results obtained, the sampling schemes affect the accuracy of the Logistic Regression and TAN classifiers, while the effect over the Naive Bayes classifier is modest. The results show a trend in imputation gain generally in favor of the variable sampling scheme. The trend is higher for TAN for which the expected gain is about 1% at the 50% missing values rate and around 2-4% at the 70% rate. And also among the imputation methods

studied, the aforementioned trend is higher for mean method.

Besides, our study shows that imputation of missing values of the two sampling schemes may result in opposing direction effects on classifier's performance. Given that the previous research has focused almost solely on the variable sampling scheme, these findings therefore provide evidence to suggest that under a fixed sampling scheme, those conclusions may not hold. We must point out that among the imputation methods considered, MIEM generally yields the least inconsistent patterns of imputation results (at higher rates of missing data) and MF the most (particularly for the LR model).

These results point to significant effects of the variable vs. fixed sampling on classifiers performance, but further investigation is required in order to understand why the advantage often goes to one scheme, and occasionally to the other.

CHAPTER 7 INFORMED SAMPLING AND IMPUTATION METHODS IN BINARY CLASSIFICATION TASKS

We finally come to the last research question (RQ5), whether the gains from informed sampling add up to the gains from imputation, or not.

7.1 Chapter Overview

In the last two decades or so, some of the substantial advances in machine learning relate to sampling techniques. For example, boosting uses resampling to improve model training, and active learning uses unlabeled data gathered so far to decide what are the most relevant data points to ask an oracle to label.

In previous work (Ghorbani and Desmarais, 2013, 2014), we established that *informed sampling*, a technique that uses feature entropy to guide the sampling, can improve the performance of some classification models. Informed sampling necessarily implies missing values, and many classifiers either require the imputation of missing values, or can often be improved by imputation. Therefore imputation and informative sampling are almost inevitably bound to be combined in practice. The obvious question is whether the gains obtained from each are additive.

In this chapter[1], two sets of experiments are performed. First, the levels of improvement from informed sampling and imputation are estimated and compared individually. Second, we investigate whether the gains obtained from each are additive by estimating the improvement obtained from informed sampling over standard uniform sampling.

The rest of the chapter is organized as follows. Below in section 7.2, our experimental methodology is explained. Then in section 7.3, the results are presented and discussed. Finally, we make some concluding remarks in section 7.4.

7.2 Methodology

The experiments have been performed using the high and low entropy informed sampling schemes, the three classifier models, the four imputation methods and the fourteen datasets introduced in previous chapters based on the procedures explained below.

[1]This study is under review at the following venue: Ghorbani, S. and Desmarais, M.C. (Under Review) Informed Sampling and Imputation Methods in Binary Classification Tasks, Journal of IEEE Transactions on Knowledge and Data Engineering.

7.2.1 Experimental Setup

A first experiment aims to assess the improvements obtained from informed sampling and imputation methods in isolation, and a second experiment aims to assess the improvements when they are combined. The methodology of each is described in the following.

First Experiment: Isolated Contributions

The first set of experiments aims to compare informed sampling with imputation. They have been carried out using the procedure illustrated in figure 7.1. For each original dataset, three samples are created such that the proportion of total missing values assigned into all attributes of the dataset is 50%:

Informed Sampling:

HE.S: High Entropy Sampling.

LE.S: Low Entropy Sampling.

Uninformed Sampling:

U.S.: Uninformed Sampling.

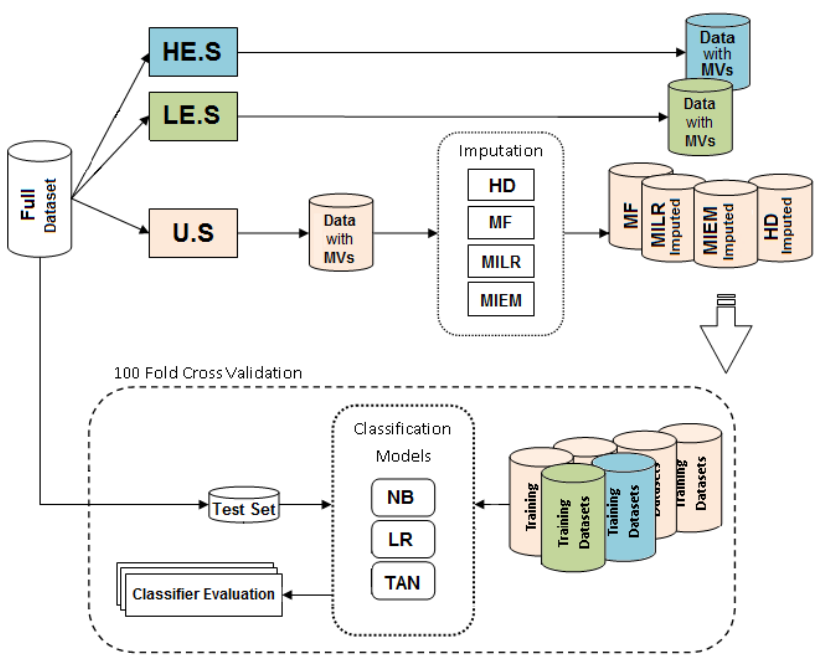

Figure 7.1 Comparison. Experimental Setup. LE.S and HE.S are respectively low and high entropy sampling schemes, whereas U.S is the uninformed, uniform sampling which is used for the imputation methods. The performance of the different classifiers over the resulting datasets is then compared

As figure 7.1 shows, the sample created using uniform scheme is subsequently imputed using the four imputation methods resulting in four imputed versions of the dataset. Note that, for LE.S and HE.S, the dataset contains missing values (not shown in the figure to avoid cluttering). If a classifier cannot handle missing values, the baseline mean imputation method is used. This applies to the LR and TAN classifiers.

Next, in a 100 fold cross validation process, the 6 samples are used with the 3 classifiers: NB, LR and TAN. Each classifier's accuracy is evaluated by applying the corresponding classification model on the test set, as shown in the figure. The same test dataset is used for all of the models over all the 6 datasets in each fold.

Second Experiment: Combined Contributions

The diagram illustrated in figure 7.2 depicts the structure of the second experiment aimed at assessing the gains from a combination of informed sampling and imputation methods. Starting from the full original dataset 3 samples are created akin to the comparison experiment.

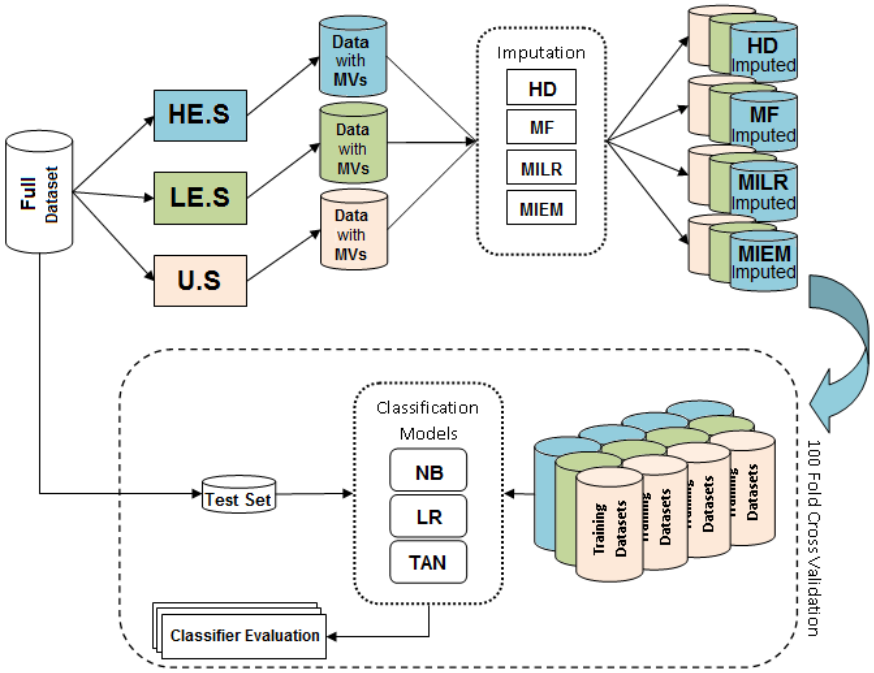

Figure 7.2 Combination Experimental Setup. The imputation methods are applied on the three samples created by HE.S, LE.S and U.S schemes. The performance of the different classifiers over the resulting datasets is then compared

The proportion of missing values assumed in this experiment is also 50%. Afterward, and in

contrast to the first experiment, all 3 samples are imputed using HD, MF, MILR and MIEM imputation methods resulting in 4 imputed versions for each sample. Next, the same cross validation process is carried out with the same classifier models using the 12 samples.

7.3 Results

Akin to our former studies, the performance of the classifiers has been assessed using a zero–one loss function in this study. Although some datasets may assume different costs for their classification decisions, we assume a uniform cost for all the classes in order to be able to compare the results across different datasets. Below, we first present the results obtained from the first set of experiments explained in section 7.2, and then those belong to the second set are demonstrated and discussed.

7.3.1 First Experiment

The first experiment aims to assess and compare the standalone gain obtained from imputation methods and from sampling schemes.

We use the gain based on the F-score:

$$F.Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7.1}$$

The F-score can be interpreted as an equally weighted average of precision and recall. In our context, precision and recall are the accuracies of the two binary classes. The best performance is at 1 and the worst at 0.

The gain is defined as the improvement in F-score over a baseline:

$$Gain_{B,i} = \frac{F.Score_i - F.Score_{base}}{1 - F.Score_{base}} \tag{7.2}$$

where $i \in \{HD, MF, MILR, MIEM, LE.S, HE.S\}$. The baseline $F.Score_{base}$ is defined by the model's performance under uniform sampling ($F.Score_{U.S}$). As mentioned earlier, when the model requires imputation, the missing values are imputed by mean imputation method for the models LR and TAN.

Figure 7.3 illustrates the average gain in error reduction, $Gain_B$, for each imputation method ($HD, MF, MILR, MIEM$) and for the two informed sampling schemes ($LE.S, HE.S$). Error bars show the confidence interval (95.0%). Averages are based on the 14 datasets and for 100 runs.

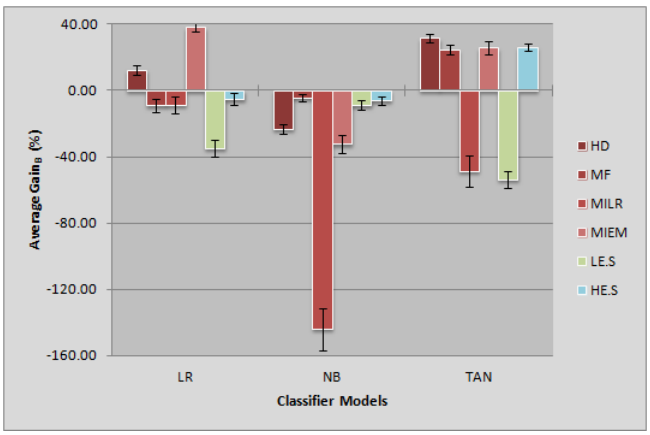

Figure 7.3 The Average gain (loss) in error reduction using formula (7.2) over 14 datasets and 100 runs. High entropy sampling can provide the TAN model with a substantial performance improvement that is comparable to imputation methods (HD, MF and MIEM). However, no gain is obtained for NB, and for the LR model two imputation methods bring substantial improvements (HD and MIEM)

The results show widely different patterns of improvements across the different classifiers.

In the case of Naive Bayes, no gain is obtained on average. The imputation methods tend to more deteriorate the performance of NB model as they appear to introduce noise rather than improve this classifier's performance. Note that these results are for 50% rate of missing values used across all experiments in this study.

For LR and TAN, some imputation methods bring substantial gains. Logistic Regression gets the highest improvements from MIEM. However, for the TAN classifier the best gains are comparable for both imputation methods and high entropy informed sampling alone, ranging between 30% and 40%.

Individual Dataset Analysis

Whilst figure 7.3 reports averages over datasets, let us look at the results in terms of individual datasets and report the percentage of datasets that show statistically significant results over the baseline, $F.Score_{base}$.

The results of a Wilcoxon signed-rank test of the difference $F.Score_i - F.Score_{base}$ is shown in figure 7.4. Orange bars represent gains and blue ones represent losses over the baseline. The results are consistent with the average gain of figure 7.3. For Naive Bayes, the large majority of datasets show losses for HD, MILR and MIEM. For Logistic Regression, the large majority show improvements for HD and MIEM, but mixed results for the other conditions. Finally, for TAN, we see statistically significant gains for 71% to 86% of the datasets for HD,

MF, MIEM and also for high entropy sampling (HE.S).

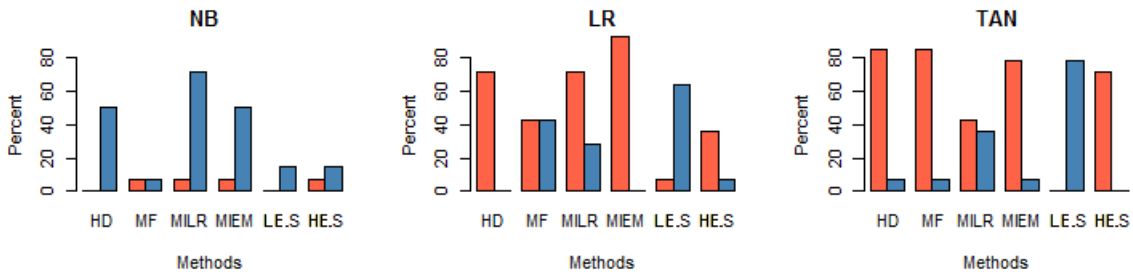

Figure 7.4 Results of Wilcoxon signed-rank test score improvements for individual datasets at $p < 0.05$. Orange bars show the percentage of datasets on which $F.Score_i > F.Score_{base}$ and the blue bars show the percentage of cases on which $F.Score_i < F.Score_{base}$ where $i \in \{HD, MF, MILR, MIEM, LE.S, HE.S\}$

7.3.2 Second Experiment

The second experiment aims to determine if the gains obtained from imputation and informed sampling in isolation are additive when the methods are combined.

A new definition of gain is defined for this purpose:

$$Gain_{I,i,s} = \frac{F.Score_{s,i} - F.Score_{U.S,i}}{1 - F.Score_{U.S,i}} \tag{7.3}$$

where $i \in \{HD, MF, MILR, MIEM\}$ and $s \in \{LE.S, HE.S\}$. $F.Score_{U.S,i}$ is the new baseline. It corresponds to the score of the uniform sampling for which the imputation method i is applied (whereas $F.Score_{base}$ corresponds to uniform sampling without imputation, which we could also refer to as $F.Score_{U.S}$). A value of $Gain_{I,i,s} = 0$ means there is no gain for an imputation method i combined to informed sampling s, over the corresponding imputation with uninformed sampling (U.S).

Figure 7.5 provides a general view of the results from the combination of informed sampling schemes and imputation methods. The figures illustrate the average $Gain_I$ for different combinations of informed sampling schemes (LE.S and HE.S) and the imputation methods. Again, error bars in the figures show the confidence interval (95.0%).

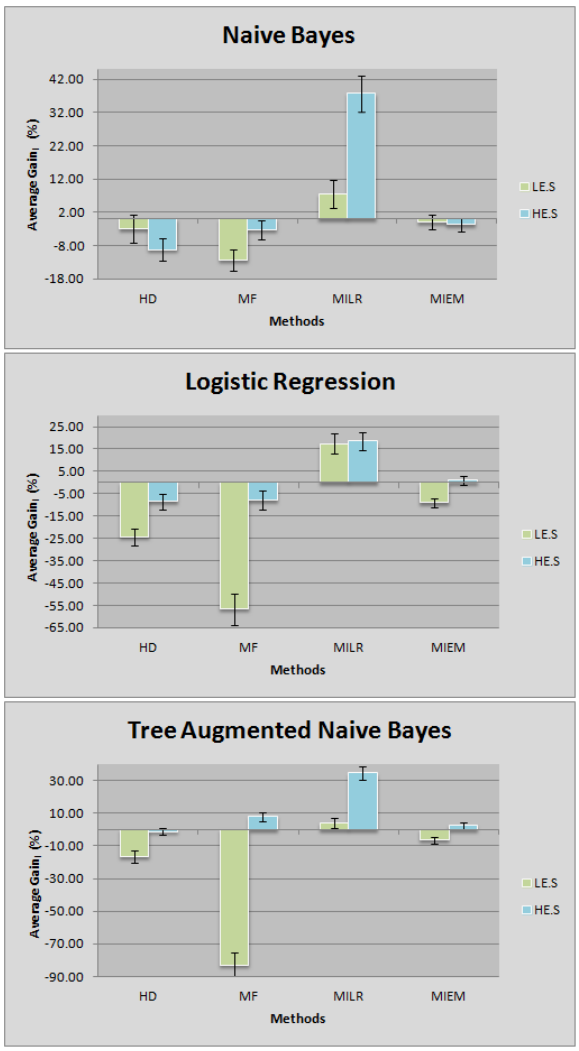

Figure 7.5 The average gain over imputation, $Gain_I$ (eq. 7.3), when the informed sampling and imputation are applied in tandem. For NB, high entropy sampling scheme with MILR results in an average $Gain_I$ of 38%. For LR, when HE.S scheme is coupled with MILR or MIEM an average $Gain_I$ of 18.5% or 1.1% can be obtained respectively. Finally, for TAN, HE.S with MF, MILR and MIEM brings the average $Gain_I$ of 8%, 35% and 2.5% respectively

We can see from figure 7.5 that substantial gains are obtained for the high entropy sampling (HE.S) condition over the uniform sampling only for the MILR imputation method. The low entropy sampling (LE.S) also brings improvement for the Logistic Regression model (LR) for the same imputation method.

The high entropy sampling also brings modest improvements (8% and 2.5%) for the TAN classifier with MF and MIEM, in addition to the substantial 35% obtained for MILR.

Individual Datasets Analysis

Akin to figure 7.4, figure 7.6 reports the number of datasets over which we observe statistically significant differences. Orange and blue bars respectively show the percent of datasets on which the classifier's $F.Score_{s,i}$ is significantly higher (orange) or lower (blue) than $F.score_{U.S,i}(p < 0.05)$.

These results are consistent with figure 7.5 and show a substantially greater number of datasets for which positive gains are obtained with high entropy and MILR combined.

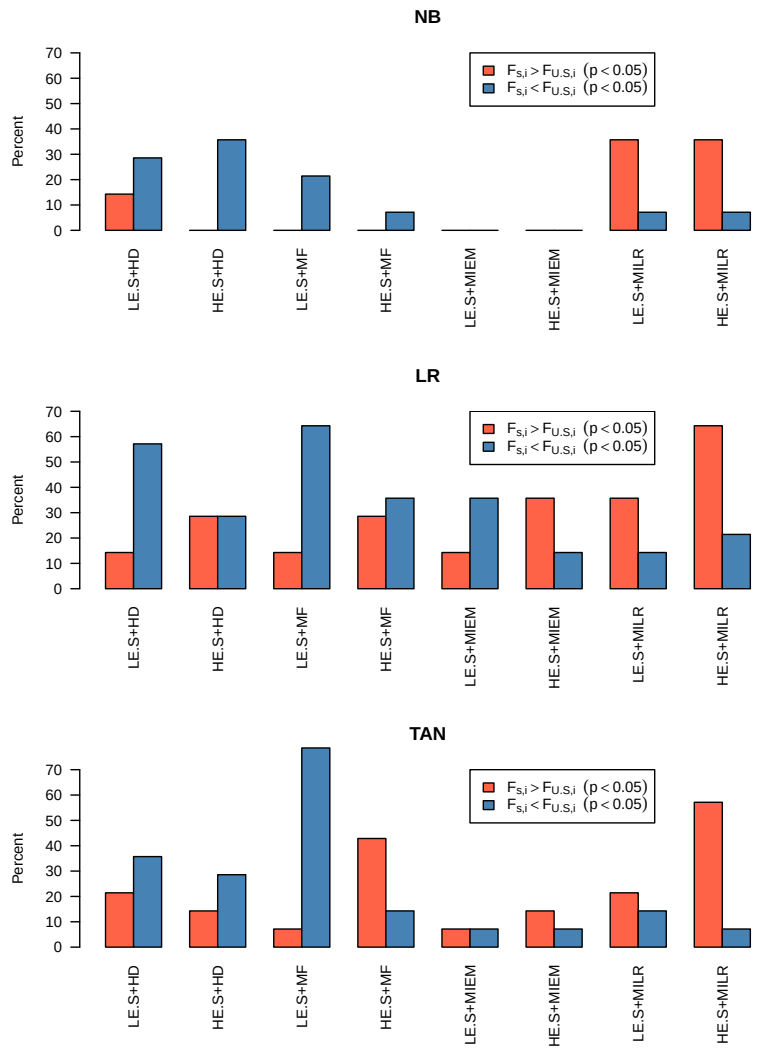

Figure 7.6 Results of Wilcoxon signed-rank test score improvements for individual datasets at $p < 0.05$. Orange bars show the percentage of datasets on which a significant difference is found for $F.Score_{s,i} > F.Score_{U.S,i}$ and the blue bars show the percentage of cases on which $F.Score_{s,i} < F.Score_{U.S,i}$ where where $i \in \{HD, MF, MILR, MIEM\}$ and $s \in \{LE.S, HE.S\}$

7.3.3 Detailed results

We conclude the presentation of the results with a final and detailed perspective of the results on a per dataset analysis, because one more question needs to be addressed: given that some combinations of imputation algorithms and informed sampling schemes can bring substantial improvements on average over each of technique alone, can a combination of imputation and sampling bring better results than the top performance we can expect without this combination?

To answer this question we have to look at the individual datasets results that are given in tables 7.1 to 7.3 and in the corresponding figures 7.7 to 7.9. Together they report the 100 runs averages of individual F-Scores per method and per dataset. The F-Scores of the figures are in fact transformed to logit units for better visualization ($1/(1+e^x)$).

The baseline $F.Score_{base}$ of equation 7.2 corresponds to the origin from where the bars are drawn (upward corresponds a gain and downward to a loss) in the figures. The black bars correspond to the individual baselines $F.Score_{U.S,i}$ for $i \in \{HD, MF, MILR, MIEM\}$. The bars in the gray area correspond to the F-Scores of informed sampling schemes without imputation.

We will not go into the details of these results, except to say that we are interested in the highest blue bars (drawn upward) and whether they are significantly higher than the black bar in their group (HD, MF, MILR, or MIEM). If they are, it means the corresponding informed sampling technique combined with the corresponding imputation method performs significantly better than any of the two techniques alone. And if this situation occurs for the best performing imputation algorithm, it implies the combination of imputation and informed sampling can bring better performance over all other non combined technique. Moreover, if the height of a blue bar corresponds to the cumulative height of the black bar in its group plus the height of the corresponding LE.S or HE.S for that dataset (in the gray area), it means the performance gains from informed sampling and imputation are truly additive.

We can see that the highest performance being attributed to blue bars occurs only for very few datasets, such as D8 and D6 for LR (figure 7.8). When looking at the best imputation method for a dataset, a black bar (a baseline) is generally on top, or on par with blue bars. Therefore, we conclude that informed sampling, whether for high or for low entropy, only rarely brings substantial improvements to the baseline of the top performing imputation methods.

7.4 Conclusion

The first experiment results show that, for the TAN classifier, high entropy informed sampling brings substantial improvements, in the same range as three imputation methods and around 30%. For Naive Bayes the large majority of datasets show losses for HD, MILR and MIEM imputation methods. And Logistic Regression gets the highest improvements from MIEM.

The second experiment addresses the question of additivity and shows that combining informed sampling with imputation brings gains with high entropy sampling scheme. On average, the significant $Gain_I$ are obtained for high entropy scheme:

- 38% for MILR with the NB model
- 18% for MILR with the LR model and,
- 8%, 35% and 2.5% for respectively MF, MILR and MIEM with the TAN model.

High entropy is the most preferable informed sampling scheme to be used in tandem with the imputation methods studied and TAN is shown to be the most receptive model to benefit from it with three of the four imputation methods considered (all but HD).

In spite of the encouraging results for some combinations of informed sampling and imputation algorithms, detailed analysis of individual dataset results reveal that these combinations rarely bring classification performance above the top imputation algorithms or informed sampling by themselves. Informed sampling often improves imputation algorithms performance, but rarely improves the best imputation algorithm(s) for a single dataset (which often is MIEM).

Table 7.1 Average F-Scores of NB classifier under different imputation/informed sampling methods over each dataset (Averaged over 100 Runs)

Dataset	Baseline	HD	LE.S+ HD	HE.S+ HD	MF	LE.S+ MF	HE.S+ MF	MILR	LE.S+ MILR	HE.S+ MILR	MIEM	LE.S+ MIEM	HE.S+ MIEM	LE.S	HE.S
D1	0.5706	0.5580	0.5020	0.5870	0.5726	0.5885	0.5819	0.5674	0.5353	0.5294	0.5556	0.5525	0.5536	0.5690	0.5504
D2	0.9568	0.9535	0.9630	0.9402	0.9557	0.9532	0.9509	0.9249	0.9239	0.9322	0.9534	0.9553	0.9513	0.9566	0.9562
D3	0.9847	0.9772	0.9817	0.9776	0.9799	0.9755	0.9818	0.9912	0.9862	0.9924	0.9929	0.9925	0.9941	0.9865	0.9887
D4	0.9877	0.9852	0.9852	0.9877	0.9867	0.9873	0.9878	0.9775	0.9786	0.9813	0.9879	0.9885	0.9867	0.9879	0.9880
D5	0.9908	0.9882	0.9912	0.9871	0.9915	0.9903	0.9917	0.9699	0.9692	0.9728	0.9908	0.9890	0.9884	0.9906	0.9920
D6	0.9839	0.9817	0.9836	0.9827	0.9840	0.9847	0.9824	0.9538	0.9611	0.9677	0.9806	0.9819	0.9821	0.9839	0.9834
D7	0.9940	0.9924	0.9888	0.9907	0.9934	0.9946	0.9942	0.9482	0.9549	0.9590	0.9810	0.9820	0.9808	0.9943	0.9936
D8	0.9849	0.9816	0.9750	0.9801	0.9866	0.9803	0.9852	0.9341	0.9336	0.9621	0.9691	0.9715	0.9707	0.9836	0.9851
D9	0.9917	0.9918	0.9888	0.9899	0.9913	0.9902	0.9938	0.9749	0.9776	0.9769	0.9860	0.9848	0.9876	0.9912	0.9926
D10	0.9932	0.9908	0.9902	0.9885	0.9943	0.9920	0.9897	0.9810	0.9865	0.9859	0.9922	0.9921	0.9917	0.9906	0.9925
D11	0.9899	0.9861	0.9888	0.9867	0.9898	0.9885	0.9885	0.9815	0.9803	0.9787	0.9902	0.9898	0.9893	0.9893	0.9893
D12	0.8822	0.8760	0.8812	0.8573	0.8874	0.8885	0.8772	0.8890	0.8848	0.8769	0.8825	0.8850	0.8761	0.8856	0.8787
D13	0.5692	0.5842	0.4955	0.5342	0.5524	0.5500	0.5790	0.5488	0.5539	0.4916	0.5803	0.5916	0.5732	0.5718	0.5308
D14	0.8960	0.8814	0.8973	0.8708	0.8859	0.8835	0.8714	0.8601	0.8354	0.8654	0.8669	0.8584	0.8652	0.8706	0.8720

Table 7.2 Average F-Scores of LR classifier under different imputation/informed sampling methods over each dataset (Averaged over 100 Runs)

Dataset	Baseline	HD	LE.S+ HD	HE.S+ HD	MF	LE.S+ MF	HE.S+ MF	MILR	LE.S+ MILR	HE.S+ MILR	MIEM	LE.S+ MIEM	HE.S+ MIEM	LE.S	HE.S
D1	0.5358	0.5679	0.4014	0.5667	0.4514	0.4733	0.4896	0.5701	0.5255	0.4944	0.5479	0.5458	0.5412	0.5150	0.5104
D2	0.9227	0.9487	0.9513	0.9340	0.9168	0.9075	0.9216	0.9157	0.9205	0.9269	0.9472	0.9511	0.9534	0.9107	0.9254
D3	0.9691	0.9778	0.9721	0.9773	0.9628	0.9092	0.9763	0.9849	0.9864	0.9873	0.9906	0.9929	0.9934	0.9659	0.9695
D4	0.9576	0.9711	0.9666	0.9746	0.9711	0.9734	0.9725	0.9773	0.9661	0.9811	0.9853	0.9852	0.9863	0.9550	0.9747
D5	0.9569	0.9647	0.9713	0.9698	0.9672	0.9738	0.9591	0.9703	0.9724	0.9762	0.9859	0.9824	0.9858	0.9594	0.9614
D6	0.9580	0.9743	0.9649	0.9663	0.9478	0.9122	0.9532	0.9617	0.9797	0.9686	0.9730	0.9762	0.9811	0.9507	0.9555
D7	0.9759	0.9846	0.9696	0.9781	0.9767	0.9665	0.9839	0.9540	0.9576	0.9626	0.9805	0.9777	0.9842	0.9653	0.9796
D8	0.9629	0.9845	0.9754	0.9618	0.9717	0.9283	0.9628	0.9395	0.9853	0.9791	0.9820	0.9802	0.9872	0.9569	0.9656
D9	0.9710	0.9758	0.9597	0.9776	0.9858	0.9822	0.9670	0.9719	0.9730	0.9869	0.9869	0.9808	0.9896	0.9535	0.9766
D10	0.9757	0.9778	0.9709	0.9765	0.9914	0.9770	0.9873	0.9809	0.9900	0.9855	0.9944	0.9897	0.9877	0.9596	0.9804
D11	0.9674	0.9710	0.9679	0.9757	0.9697	0.9443	0.9707	0.9804	0.9850	0.9747	0.9890	0.9821	0.9893	0.9770	0.9764
D12	0.8940	0.8980	0.9090	0.9092	0.8715	0.8277	0.8095	0.9181	0.9129	0.8770	0.9213	0.9212	0.9143	0.8746	0.8833
D13	0.5402	0.5001	0.4201	0.5236	0.5258	0.5158	0.5065	0.4462	0.5244	0.5595	0.5905	0.5384	0.5240	0.5374	0.5234
D14	0.8752	0.8758	0.8586	0.8897	0.8362	0.8588	0.8588	0.9240	0.8986	0.9087	0.9219	0.9115	0.9155	0.8572	0.8850

Table 7.3 Average F-Scores of TAN classifier under different imputation/informed sampling methods over each dataset (Averaged over 100 Runs)

Dataset	Baseline	HD	LE.S+ HD	HE.S+ HD	MF	LE.S+ MF	HE.S+ MF	MILR	LE.S+ MILR	HE.S+ MILR	MIEM	LE.S+ MIEM	HE.S+ MIEM	LE.S	HE.S
D1	0.5076	0.5390	0.4743	0.5397	0.4569	0.5112	0.5156	0.5763	0.5404	0.5687	0.5683	0.5635	0.5341	0.4981	0.5030
D2	0.9246	0.9496	0.9575	0.9435	0.9460	0.9033	0.9403	0.9246	0.9225	0.9362	0.9551	0.9537	0.9519	0.9064	0.9291
D3	0.9557	0.9860	0.9901	0.9851	0.9721	0.9511	0.9824	0.9915	0.9875	0.9935	0.9948	0.9940	0.9944	0.9325	0.9738
D4	0.9678	0.9838	0.9844	0.9862	0.9845	0.9688	0.9877	0.9747	0.9806	0.9845	0.9879	0.9881	0.9872	0.9564	0.9833
D5	0.9790	0.9878	0.9873	0.9902	0.9867	0.9804	0.9886	0.9706	0.9689	0.9773	0.9908	0.9886	0.9879	0.9595	0.9886
D6	0.9678	0.9822	0.9810	0.9838	0.9825	0.9290	0.9808	0.9593	0.9567	0.9692	0.9797	0.9815	0.9809	0.9567	0.9826
D7	0.9897	0.9924	0.9823	0.9901	0.9885	0.9528	0.9917	0.9473	0.9548	0.9657	0.9848	0.9794	0.9854	0.9883	0.9923
D8	0.9703	0.9836	0.9713	0.9810	0.9838	0.9499	0.9863	0.9371	0.9385	0.9688	0.9715	0.9756	0.9773	0.9438	0.9839
D9	0.9883	0.9906	0.9878	0.9914	0.9881	0.9898	0.9905	0.9758	0.9770	0.9808	0.9850	0.9839	0.9884	0.9833	0.9920
D10	0.9811	0.9920	0.9915	0.9910	0.9914	0.9706	0.9888	0.9807	0.9883	0.9853	0.9924	0.9916	0.9926	0.9593	0.9890
D11	0.9772	0.9900	0.9895	0.9878	0.9841	0.9681	0.9878	0.9808	0.9793	0.9800	0.9914	0.9893	0.9909	0.9700	0.9836
D12	0.8730	0.8869	0.8969	0.9024	0.9032	0.8795	0.8939	0.9113	0.9111	0.8870	0.9224	0.9154	0.9264	0.8434	0.8834
D13	0.5291	0.4889	0.4618	0.4516	0.5535	0.5280	0.5695	0.4915	0.5203	0.5176	0.5841	0.5643	0.5815	0.5007	0.5464
D14	0.8322	0.8905	0.8530	0.8953	0.8647	0.8185	0.8665	0.9256	0.8733	0.9156	0.9097	0.9094	0.9156	0.8035	0.8546

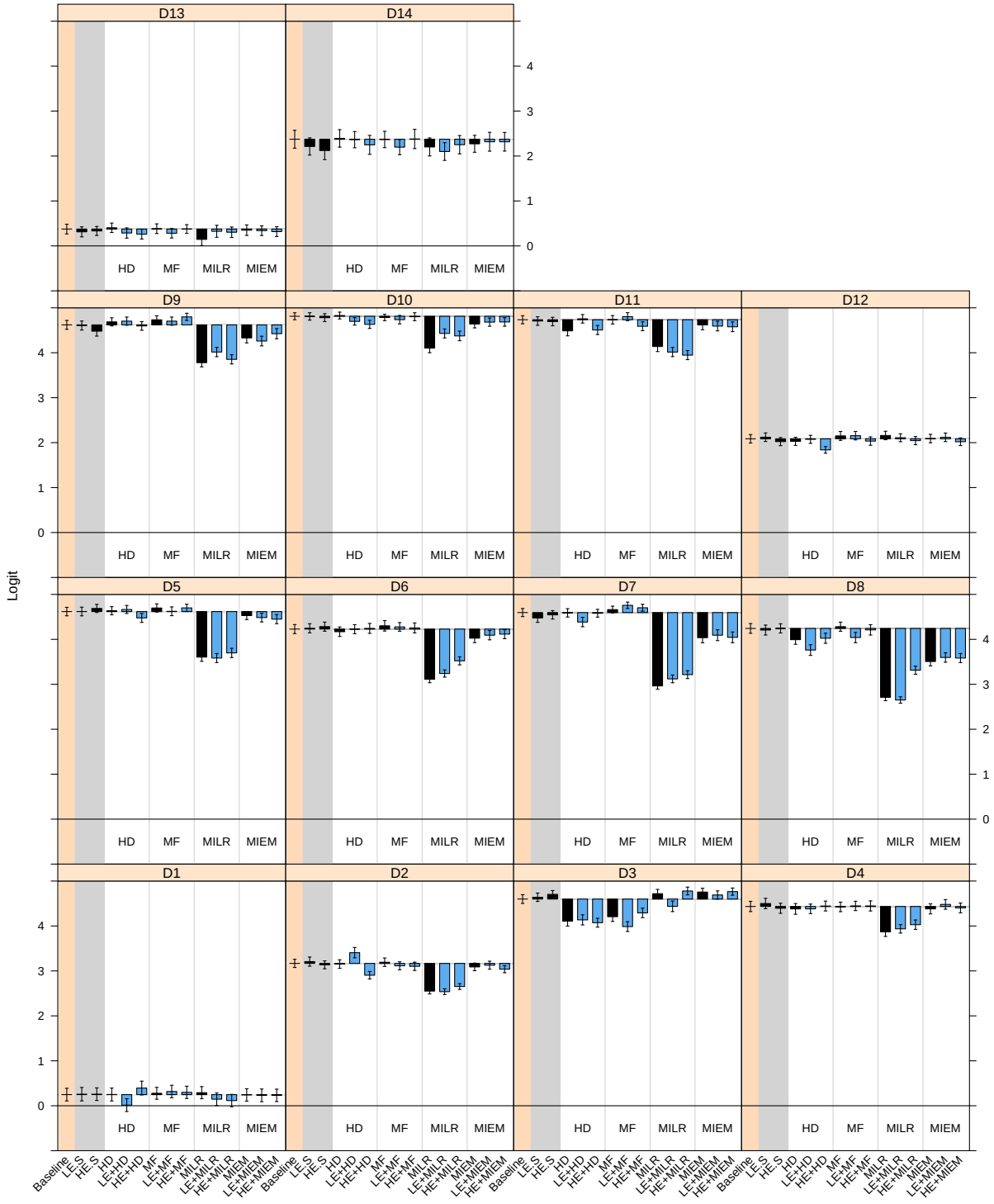

Figure 7.7 Individual dataset logit of F-scores for NB. Bars are shown relative to the baseline

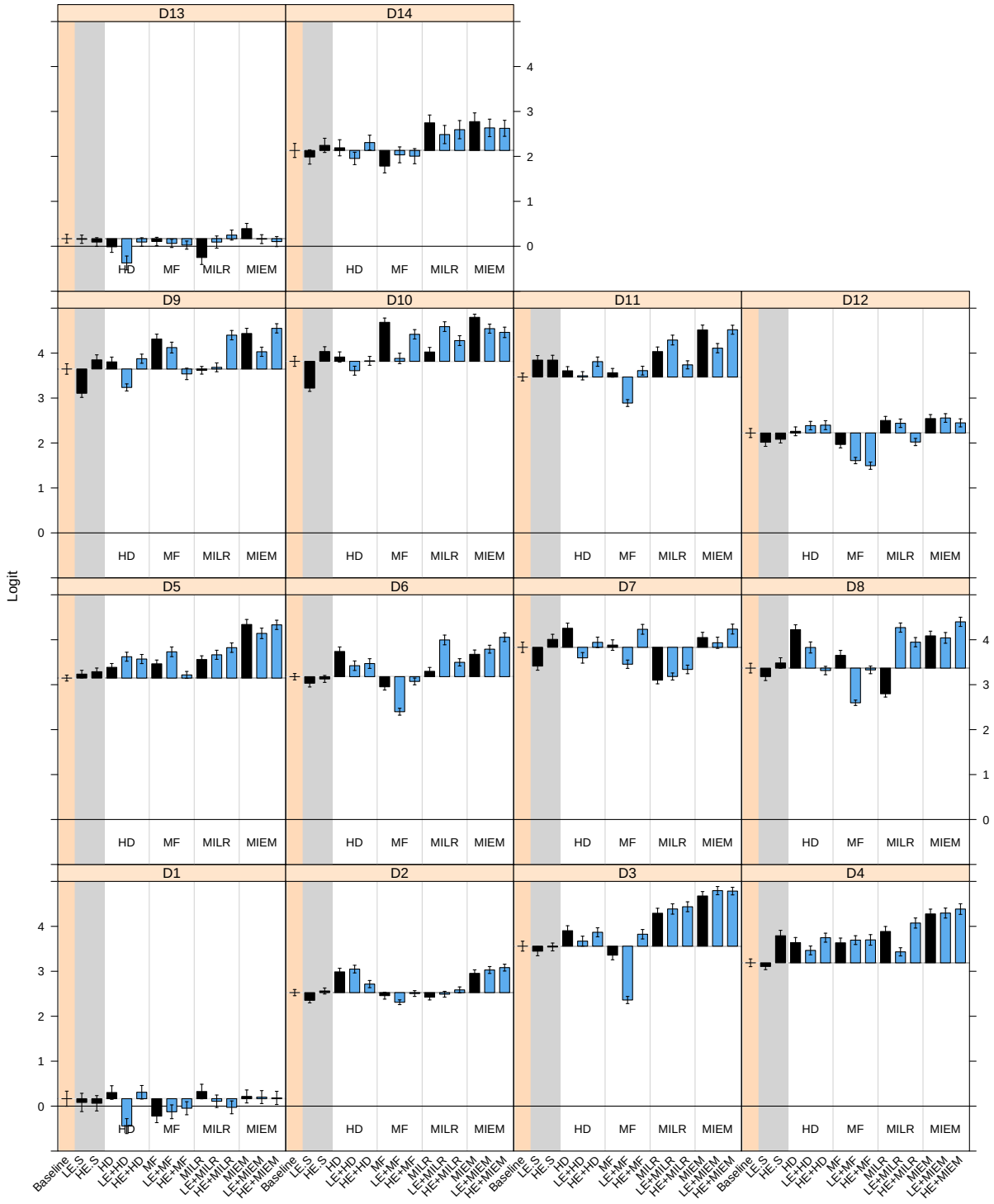

Figure 7.8 Individual dataset logit of F-scores for LR. Bars are shown relative to the baseline

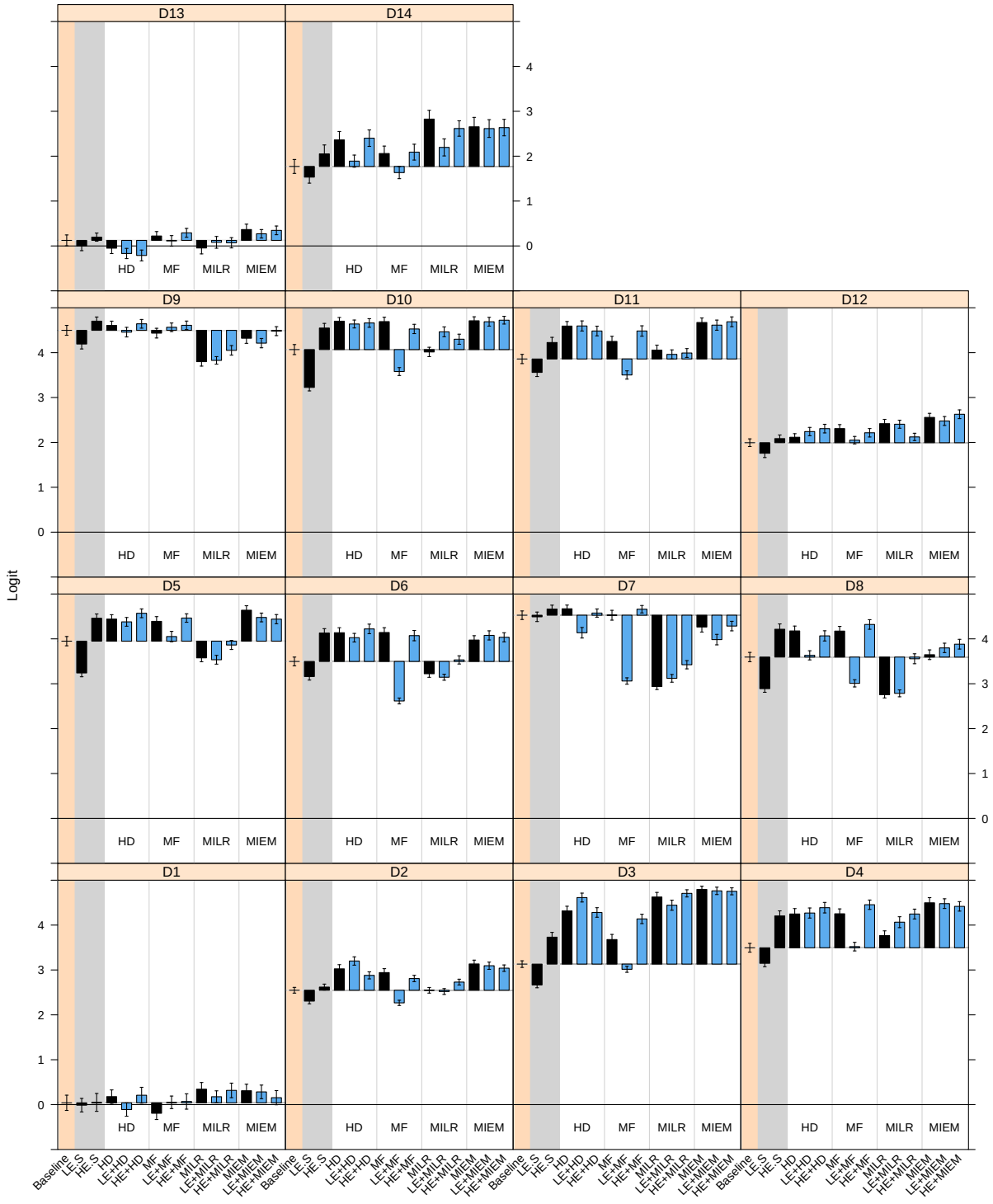

Figure 7.9 Individual dataset logit of F-scores for TAN. Bars are shown relative to the baseline

CHAPTER 8 CONCLUSION AND RESEARCH PERSPECTIVES

8.1 Conclusion

The study was set out to investigate the situation where a fixed number of observations is allocated per case record, and where the choice is given among which variables to sample from during the data gathering phase. One main objective of the thesis was to study how the planning of missing data, by conducting an informed sampling process which relies on features entropy, helps improve the performance of classification methods. We focused on three classifier models, namely Logistic regression, Naive Bayes, and Tree Augmented Naive Bayes over a binary classification task and formalized the objective through two concise research questions as follows:

- **RQ1**- Can a selective sampling approach based on uncertainty improve the performance of classifiers?

  In the study presented in Chapter 3, we defined three different schemes of sampling: 1-Uniform random sampling as a baseline, 2-Low entropy sampling (greater sampling rate for low entropy variables) and 3-High entropy sampling (greater sampling rate for higher entropy variables) and assessed the classification performance of the models under the different sampling schemes. The results showed that the third scheme systematically improves the prediction performance of the TAN classifier. However, for the NB and LR classifiers the improvements were obtained for only half of the datasets, mostly from high-entropy sampling and for one dataset from low-entropy scheme.

- **RQ2**- Can we guide the selective sampling approach on the fly, during the data gathering phase?

  The results of the former study were obtained given a priori information on the entropy of each variable. To validate that classification improvements can be obtained in a realistic setting where such information is generally not available, we developed an adaptive algorithm to guide the sampling heuristic (Chapter 4), where each variable's entropy is measured with seed data and updated on the fly. Results showed that using a small seed dataset, the improvements obtained from the former study hold for TAN and LR.

Selective sampling necessarily involves missing values, which in turns generally involves imputation of missing values, since a number of classifiers either require or benefit from imputation. We therefore addressed the question of how the performance gains from selective sampling combine with imputation techniques. We first explore the performance of different imputation techniques under different conditions.

- **RQ3**- Can imputation techniques improve the prediction accuracy of classification tasks with a fixed rate of observations per record?

  Previous research has solely relied on non-fixed (variable) rate of observations per record and for a very limited scope of relatively small amounts of missing values. In a comprehensive study presented in Chapter 5, we assessed the predictive accuracy improvements obtained from two single imputation methods, missForest and Sequential Hot-Deck, and two Multiple Imputation methods, one based on Expectation-Maximization, and the other based on Logistic Regression across a wide range of missing data rates. The results showed that as in previous studies with non-fixed observations per record, imputation techniques improve the prediction accuracy of different classification models and that the patterns of improvement brought by imputation algorithms can vary substantially per model and also per missing value rate. In addition, the MIEM method was shown to generally give the best results for the classifiers considered across almost all rates of missing data. The analysis of the results also showed that TAN and particularly, LR receive the most classification performance gains obtained from the imputation.

We also investigated the improvements along a different dimension to study whether the choice of the rate of sampling per record between fixed and variable affects the performance of imputation methods, as stated by the following research question:

- **RQ4**- What is the impact of variable vs. fixed sampling rate per record on the performance of imputation methods in classification tasks?

  To address the question, in another study elaborated in Chapter 6, we showed that despite the same total number of observations, the two sampling schemes do affect the classifier performance, but the effect varies between models and missing values rates. Analysis of the obtained results showed that, compared to the fixed sampling scheme, variable scheme generally results in higher prediction performance of classification when is used in tandem with imputation and that the TAN classifier and Mean imputation method are the most affected by the variable scheme.

In light of the aforementioned studies, classifier improvement gains are expected to be obtained from both informed sampling and imputation. Since imputation and informed sam-

pling are likely to be combined in practice, the obvious question is:

- **RQ5**- Are the improvements from informed sampling and from imputation additive?

  In a final study presented in Chapter 7, we first compared the levels of improvement from informed sampling with those from a number of imputation techniques. Next, we investigated whether the gains obtained from sampling and imputation are additive. The results showed that the individual gains from informed sampling and imputation are within the same range and that combining high-entropy informed sampling with imputation brings significant gains to the classifiers' performance. TAN was shown to be the most receptive model to benefit from informed sampling with most of the imputation methods considered. And generally, the gains are not additive. Detailed analysis of individual dataset results also revealed that the combinations rarely bring classification performance above the top imputation algorithms or informed sampling by themselves. Informed sampling often improves imputation algorithms performance, but rarely improves the best imputation algorithm(s) for a single dataset (which often is MIEM).

In summary, the main contributions of this dissertation included:

1. An entropy-based selective sampling design to improve the performance of classification models
2. A demonstration that the gains from the proposed informed sampling approach are maintained when entropy is estimated on the fly
3. An extensive assessment of the impact of imputation methods under different sampling schemes that resulted in the following demonstrations:
   - Imputation can improve the performance of classifiers over data with fixed rate and high proportions of missing values per record
   - The choice of the rate of sampling per record between fixed and variable affects the performance of imputation methods
4. A comprehensive study on the potential additive effects of informed sampling and imputation methods on classification performance resulted in:
   - Showing that individual gains from informed sampling and imputation are within the same range

- Demonstrating that the informed sampling often improves imputation algorithms performance, but in general the gains from the two are not additive

8.2 Limitations and Threats to Validity

In this section we list some limitations of the studies performed in this thesis and discuss some potential validity threats.

8.2.1 Generalizability

In the research presented in this dissertation we showed that given a fixed budget of observations per case record, applying our informed sampling approach can improve the prediction performance of some classification models. In the same context (even when the missing data proportion is high), we demonstrated that not only imputation techniques can be helpful as well, but also the combination of the two can bring higher levels of accuracy to the classifiers. We also showed that compared to the fixed the variable sampling generally results in higher classification performance when coupled with imputation. One important question facing this research is how the conclusions we made in this thesis generalize. In the following we discuss this in more details.

1. **Does it generalize to non-binary variables?**

   In this thesis we limited our scope to binary target variables and binary attributes. The promising results of applying the high entropy sampling on binary data can be a good reason to be optimistic about the effectiveness of informed sampling over at least nominal attributes as well (We note that the binary is indeed a sub-type of nominal scale). This can be investigated as an extension to our research.

   For the imputation, given that numerous studies have shown the effectiveness of imputation to improve classification accuracy with almost all data types on non-fixed observations per record, it is highly likely that the conclusions we made on the fixed observations of binary data will generalize to non-binary data as well. This also can be verified with a new study.

   Besides, we think that the variable sampling scheme retains much more informative records (i.e., records presenting the most relevant features to train the classifier) compared to those of the fixed sampling. This explains why the results of our study were generally in favor of the variable scheme. Based on this, we expect our conclusion on

generally higher performance of the variable scheme generalize to non-binary variables as well. This can be another interesting extension to our work.

2. **Does it generalize to other classification models?**

   In this research we focused on three classification models, among which only for TAN the improvements obtained from the informed sampling are substantial and systematic. One explanation could be that the model is more sensitive to the variables correlation than LR and NB. This might be a factor that affects how the conclusions generalize to other classifiers. This for sure, needs to be more studied. Furthermore, whether and how the informed sampling makes difference to the bias and variance errors is something we currently don't have an explanation for. This might be another factor that impacts the generalizability of our conclusions and needs to be further studied.

   But again, based on the same argument we presented in response to the former question, our conclusions on the effects of imputation are likely to generalize to other classifiers as well.

3. **Does it generalize to regression tasks as well?**

   This thesis shed light on the potential improvements the informed sampling approach can bring to classification performance. The investigation of whether and how it might be helpful with regression tasks can be another interesting area of future research.

8.3 Future Work

As alluded in the previous section, the research in this thesis points to a number of promising directions for future work, which will complement this dissertation. In addition to the studies proposed, some issues and directions warrant consideration:

- **More Effective Guidance of Informed Sampling in Practice:**

  Future work should focus on bringing a better understanding of the mechanisms at play in order to guide the use of informed sampling in practice. As discussed in the previous section, further analysis and investigations are required to determine what the best sampling scheme is in the given context. To achieve this goal we plan to perform the following research activity:

  - **Using Synthetic Data:** This approach can help us better understand the nature of interactions between the different parameters including classifier model, sampling scheme, imputation method, missing rate and, dataset characteristics. This

way the data are generated with all sorts of important properties (independence, etc) and as large as desired. Then by testing the different imputation methods and sampling schemes on these different versions of the data and keeping track of how the performance of the models we test changes as the properties of the data change, we can try to achieve the goal.

- **Better Understanding of the Effects of Variable vs. Fixed Sampling Rates per Record on Imputation methods:**

  In section 8.2 we discussed one possible explanation on why the variable sampling brings higher levels of accuracy to the classification models compared to the fixed sampling. We plan to further investigate this through Feature Selection and possibly through the same strategy explained above: Using Synthetic Data.

- **More Efficient Choice of Imputation Methods in Practice:**

  According to previous research (which as mentioned earlier, relies on variable sampling rate per record), there is no universally best imputation method that performs best for all classifiers and missing data proportions. As our study on fixed sampling rate (Chapter 5) showed, in spite of superior performance of MIEM across different classifiers and missing values rate, the patterns of effects of imputation methods can vary substantially per model and also per missing values rate. Further analysis and investigations are required to determine what the best imputation method is in a given context. As one approach to achieve this goal, we plan to conduct the following research activity:

  – **To Predict the Best Method for a Given Context:** This way we train a classifier using the features extracted and collected from our previous experiments. The training data include important items, such as classifier model, sampling scheme (fixed or variable rate per record), missing rate, dataset characteristics and the best performer in each case (class). Then, depending on the conditions in practice, it might help us find the best imputation method to apply. Replicating our previous experiments over reduced versions of the datasets (reduced sample size and or features numbers) provides us with much more data available to be used for this purpose.

REFERENCES

J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," Journal of Multiple-Valued Logic and Soft Computing, 2010.

C. Alzola and F. Harrell, "An introduction to s-plus and the hmisc and design libraries," 1999.

L. C. Anigbo, "Demonstration of the multiple matrices sampling technique in establishing the psychometric characteristics of large samples," Journal of Education and Practice, vol. 2, no. 3, pp. 19–25, 2011.

K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

F. B. Baker and S.-H. Kim, Item Response Theory, Parameter Estimation Techniques (2nd ed.). New York, NY: Marcel Dekker Inc., 2004.

A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," Journal of School Psychology, vol. 48, no. 1, pp. 5–37, 2010.

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.

E. B. Baum and K. Lang, "Query learning can work poorly when a human oracle is used," in International Joint Conference on Neural Networks, vol. 8, 1992.

L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and regression trees. CRC press, 1984.

G. Casella and E. I. George, "Explaining the gibbs sampler," The American Statistician, vol. 46, no. 3, pp. 167–174, 1992.

I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 150–157.

A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 37, no. 5, pp. 692–709, 2007.

A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," Pattern Recognition, vol. 41, no. 12, pp. 3692 – 3705, 2008.

N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine learning, vol. 29, no. 2-3, pp. 131–163, 1997.

A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka, "Selective sampling for example-based word sense disambiguation," Computational Linguistics, vol. 24, no. 4, pp. 573–597, 1998.

P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," Neural Computing and Applications, vol. 19, no. 2, pp. 263–282, 2010.

I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," Neurocomputing, vol. 73, no. 16, pp. 3039–3065, 2010.

S. Ghorbani and M. C. Desmarais, "Selective sampling designs to improve the performance of classification methods," in Machine Learning and Applications (ICMLA), 2013 12th International Conference on, vol. 2. IEEE, 2013, pp. 178–181.

——, "An adaptive sampling algorithm to improve the performance of classification models," in European Conference on Data Mining 2014, vol. 1. IADIS, 2014, pp. 21–28.

N. Golbandi, Y. Koren, and R. Lempel, "On bootstrapping recommender systems," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 1805–1808. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871734

J. W. Graham, "Missing data analysis: Making it work in the real world," Annual review of psychology, vol. 60, pp. 549–576, 2009.

J. W. Graham, S. M. Hofer, and D. P. MacKinnon, "Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures," Multivariate Behavioral Research, vol. 31, no. 2, pp. 197–218, 1996.

J. W. Graham, B. J. Taylor, and P. E. Cumsille, "Planned missing-data designs in analysis of change." 2001.

J. W. Graham, B. J. Taylor, A. E. Olchowski, and P. E. Cumsille, "Planned missing data designs in psychological research." Psychological methods, vol. 11, no. 4, p. 323, 2006.

J. Honaker, G. King, M. Blackwell et al., "Amelia ii: A program for missing data," Journal of Statistical Software, vol. 45, no. 7, pp. 1–47, 2011.

K. Hornik, C. Buchta, and A. Zeileis, "Open-source machine learning: R meets Weka," Computational Statistics, vol. 24, no. 2, pp. 225–232, 2009.

R. D. King, K. E. Whelan, F. M. Jones, P. G. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," Nature, vol. 427, no. 6971, pp. 247–252, 2004.

R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova et al., "The automation of science," Science, vol. 324, no. 5923, pp. 85–89, 2009.

V. Krishnamurthy, "Algorithms for optimal scheduling and management of hidden markov model sensors," Signal Processing, IEEE Transactions on, vol. 50, no. 6, pp. 1382–1397, 2002.

D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc., 1994, pp. 3–12.

K.-H. Li, "Imputation using markov chains," Journal of Statistical Computation and Simulation, vol. 30, no. 1, pp. 57–79, 1988.

R. Little and D. Rubin, Statistical Analysis With Missing Data, ser. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1987. [Online]. Available: http://books.google.ca/books?id=w40QAQAAIAAJ

P. Liu and L. Lei, "Missing data treatment methods and nbi model," in Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on, vol. 1. IEEE, 2006, pp. 633–638.

J. Luengo, S. García, and F. Herrera, "A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between rbfns and eventcovering method," Neural Networks, vol. 23, no. 3, pp. 406–418, 2010.

——, "On the choice of the best imputation methods for missing values considering three groups of classification methods," Knowledge and information systems, vol. 32, no. 1, pp. 77–108, 2012.

D. J. MacKay, Information theory, inference and learning algorithms. Cambridge university press, 2003.

E. T. Matsubara, R. C. Prati, G. E. Batista, and M. C. Monard, "Missing value imputation using a semi-supervised rank aggregation approach," in Advances in Artificial Intelligence-SBIA 2008. Springer, 2008, pp. 217–226.

R. F. Palmer and D. R. Royall, "Missing data? plan on it!" Journal of the American Geriatrics Society, vol. 58, no. s2, pp. S343–S348, 2010.

J. Quinlan, "C4. 5: Program for machine learning morgan kaufmann," San Mateo, CA, USA, 1993.

A. Robitzsch, T. Kiefer, A. George, A. Uenlue, and M. Robitzsch, "Package CDM," 2012. [Online]. Available: http://cran.r-project.org/web/packages/CDM/index.html

D. B. Rubin, "Multiple imputation for nonresponse in surveys. new york: J," 1987.

D. B. Rubin and J. L. Schafer, "Efficiently creating multiple imputations for incomplete multivariate normal data," in Proceedings of the Statistical Computing Section of the American Statistical Association, vol. 83, 1990, p. 88.

I. G. Sande, "Hot-deck imputation procedures," Incomplete data in sample surveys, vol. 3, pp. 334–350, 1983.

B. Settles, "Active learning literature survey," University of Wisconsin, Madison, vol. 52, no. 55-66, p. 11, 2010.

——, "Active learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, no. 1, pp. 1–114, 2012.

B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008, pp. 1070–1079.

D. M. Shoemaker, "Principles and procedures of multiple matrix sampling." 1971.

Q. Song, M. Shepperd, X. Chen, and J. Liu, "Can k-nn imputation improve the performance of c4. 5 with small software project data sets? a comparative evaluation," Journal of Systems and Software, vol. 81, no. 12, pp. 2361–2370, 2008.

D. J. Stekhoven and P. Buhlmann, "Missforest-non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.

M. Templ, A. Alfons, A. Kowarik, and B. Prantner, "Vim: visualization and imputation of missing values," R package version, vol. 2, no. 3, 2011.

G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," Speech Communication, vol. 45, no. 2, pp. 171–186, 2005.

B. Twala, "An empirical comparison of techniques for handling incomplete data using decision trees," Applied Artificial Intelligence, vol. 23, no. 5, pp. 373–405, 2009.

S. Van Buuren and K. Oudshoorn, "Flexible multivariate imputation by mice," Leiden, The Netherlands: TNO Prevention Center, 1999.

V. Vapnik, "Book review: the nature of statistical learning theory," Technometrics, vol. 38, no. 4, pp. 400 – 400, 1996.

J. Vomlel, "Evidence propagation in Bayesian networks for computerized adaptive testing," vol. 12, 2002.

I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

H. Yu, "Svm selective sampling for ranking with application to data retrieval," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005, pp. 354–363.

C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," Multimedia, IEEE Transactions on, vol. 4, no. 2, pp. 260–268, 2002.

APPENDIX A Some More Analysis

Student's T-Test

In addition to the box plot representation illustrated in figure 5.5, we analyze the statistical significance of differences in prediction accuracy between using the imputation methods and directly applying the data with missing values based on paired t-tests at the 95% significance level. We use the paired t-test in which each member of one numerical set is assumed to have a unique relationship with a particular member of the other set. Tables A.1 and A.2 report the results of two different t-test approaches.

Table A.1 Datasets on which the mean accuracies of the classifiers (over 10 runs and 6 missing rates) on imputed and non-imputed data are significantly different at $p < 0.05$.

	Imputation Methods				
Classifiers	**HD**	**MF**	**MILR**	**MIEM**	**Full Dataset**
NB	D3	-	D2, D4, D5, D6, D7, D8, D9, D10, D11, D13	D7, D8, D9	-
LR	D2, D4, D6, D8	D7, D10, D14	D3, D4, D5, D6, D7, D10, D11, D13	D2, D3, D4, D5, D6, D7, D8, D10, D11, D13	D2, D3, D4, D5, D8, D10, D11
TAN	D2, D3, D4, D5, D6, D10, D11	D2, D3, D4, D5, D6, D8, D10, D12	D1, D3, D7, D8, D9, D12, D13, D14	D2, D3, D4, D5, D6, D9, D10, D11, D12, D13	D2, D3, D4, D5, D6, D8, D10, D11, D12, D13, D14
SVM	D2, D5, D6, D7, D8	D4, D5, D9, D10, D12, D13	D1, D2, D3, D6, D7, D8, D10, D11, D12, D13, D14	D1, D4, D5, D6, D8, D9, D10, D11, D12, D13	D1, D4, D5, D9, D10, D12

Table A.2 Datasets on which the mean accuracies of the classifiers (over 10 runs) on imputed and non-imputed data are significantly different at $p < 0.05$.

		Imputation Methods				
Classifiers	**MR**	**HD**	**MF**	**MILR**	**MIEM**	**Full Dataset**
LR	5%	-	-	D6, D8, D11, D13	D2, D7, D8, D11	-
	10%	-	D8, D10	D6, D8, D11	D4, D6, D8	D8
	20%	D2, D7	D8, D10	D3, D4, D6, D11, D13	D3, D4, D5, D6, D8, D10, D11, D13	D3
	30%	D2, D6, D7, D10	D10, D11, D12	D5, D6, D7, D9, D13, D14	D2, D3, D4, D5, D6, D8, D11, D13	D4
	40%	D6, D8, D10, D12, D13	D7, D8, D9, D10, D12	D2, D4, D7, D10, D12, D13	D4, D5, D6, D8, D9, D10, D11, D12, D13	D8, D9, D10, D12
	50%	D2, D6, D8	D4, D9, D10, D14	D3, D4, D7, D8, D11, D13, D14	D2, D3, D4, D5, D6, D8, D10, D11, D13	D2, D3, D4, D5, D10, D11
NB	5%	-	-	-	-	-
	10%	-	-	D9	-	-
	20%	-	-	D5, D7, D8, D9, D13	-	-
	30%	D3	-	D4, D5, D6, D7, D8, D9, D10, D13	D7, D8, D9	-
	40%	D3	-	D2, D5, D6, D7, D8, D9, D11	-	-
	50%	-	-	D2, D4, D5, D6, D7, D8, D9, D10	D7, D8, D9	D7, D9
TAN	5%	-	-	D3, D9	D3	-
	10%	-	D3	D3	D3, D4	D3
	20%	D3, D11	D3, D11	D3, D7, D8, D9, D13	D3, D5, D9, D11	D4
	30%	D2, D3, D5, D11	D3, D8, D13	D1, D3, D6, D7, D8, D9, D13	D1, D2, D3, D5, D11, D13	D2, D3, D5, D8, D11, D13
	40%	D3, D5, D6, D10, D11	D3, D4, D5, D6, D10, D11, D12	D2, D3, D7, D8, D9, D12, D13	D1, D3, D5, D6, D10, D11, D12	D3, D5, D6, D8, D10, D11, D12
	50%	D1, D2, D3, D4, D6, D8, D10, D11, D13	D2, D4, D6, D8, D10, D12	D3, D4, D5, D7, D8, D9, D12, D13	D2, D3, D4, D5, D10, D11, D12, D13	D2, D3, D4, D5, D6, D8, D9, D10, D11, D12, D13
SVM	5%	-	-	D10, D11, D14	D2, D4, D7, D9, D10, D11	-
	10%	-	-	D10, D13	D7, D10, D11, D13	-
	20%	-	D9	D7, D8	D9, D11	D9
	30%	D8	-	D2, D6, D7, D8, D13	D4, D9, D10, D13	D9
	40%	D2, D4, D5, D6, D7	D2, D9, D10, D12	D2, D6, D7, D8, D12, D13	D1, D2, D6, D10, D11, D12, D13	D10, D12
	50%	D1, D3, D7	D4, D5	D2, D6, D7, D8, D13	D4, D5, D10, D13	D4, D5

On What Datasets Applying the Imputation Methods Does not Improve the Accuracy?

Table A.3 lists the datasets on which none of the considered imputation methods can bring an improved prediction accuracy to the classifiers studied at each missing rate.

Table A.3 Datasets on which none of the imputation methods improves the classification accuracy.

	Classifiers			
Missing Rates	**NB**	**LR**	**TAN**	**SVM**
5%	D1, D2, D4	D1, D9	D10	D6
10%	D7, D10, D11	D12	D9, D11, D14	D6, D8
20%	D1, D2, D6, D7, D8, D11	D12	D1, D6, D8, D9	D3, D6, D8
30%	D6, D7, D10	D9	-	D2, D6, D8
40%	D3, D11	-	-	D2, D3, D6, D7, D8
50%	D6, D8, D11	-	-	D2, D6, D7, D8