IntechOpen

IntechOpen Series
Artificial Intelligence, Volume 39

Association Rule Mining and Data Mining

Recent Advances, New Perspectives and Applications

Edited by Jainath Yadav

Association Rule Mining and Data Mining - Recent Advances, New Perspectives and Applications

Edited by Jainath Yadav

Published in London, United Kingdom

Association Rule Mining and Data Mining - Recent Advances, New Perspectives and Applications
http://dx.doi.org/10.5772/intechopen.1001745
Edited by Jainath Yadav

Contributors
Aditya Kumar, Jainath Yadav, Kagisho Madikadike Molabe, Kwena Mokoena, Madikadike Kagisho Molabe, Madumetja Cyril Mathapo, Michal Koren, Or Peretz, Rankotsane Victoria Hlokoe, Thobela Louis Tyasi, Victoria Rankotsane Hlokoe

First published in London, United Kingdom, 2025 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

For EU product safety concerns: IN TECH d.o.o., Prolaz Marije Krucifikse Kozulić 3, 51000 Rijeka, Croatia, info@intechopen.com or visit our website at intechopen.com.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Association Rule Mining and Data Mining - Recent Advances, New Perspectives and Applications
Edited by Jainath Yadav
p. cm.

This title is part of the Artificial Intelligence Book Series, Volume 39
Topic: Machine Learning and Data Mining
Series Editor: Andries Engelbrecht
Topic Editor: Marco Antonio Aceves-Fernández

Print ISBN 978-0-85466-638-6
Online ISBN 978-0-85466-637-9
eBook (PDF) ISBN 978-0-85466-639-3
ISSN 2633-1403

IntechOpen Book Series

Artificial Intelligence

Volume 39

Aims and Scope of the Series

Artificial Intelligence (AI) is a rapidly developing multidisciplinary research area that aims to solve increasingly complex problems. In today's highly integrated world, AI promises to become a robust and powerful means for obtaining solutions to previously unsolvable problems. This Series is intended for researchers and students alike interested in this fascinating field and its many applications.

Meet the Series Editor

Andries Engelbrecht received the Masters and Ph.D. degrees in Computer Science from the University of Stellenbosch, South Africa, in 1994 and 1999 respectively. He is currently appointed as the Voigt Chair in Data Science in the Department of Industrial Engineering, with a joint appointment as Professor in the Computer Science Division, Stellenbosch University. Prior to his appointment at Stellenbosch University, he has been at the University of Pretoria, Department of Computer Science (1998-2018), where he was appointed as South Africa Research Chair in Artifical Intelligence (2007-2018), the head of the Department of Computer Science (2008-2017), and Director of the Institute for Big Data and Data Science (2017-2018). In addition to a number of research articles, he has written two books, Computational Intelligence: An Introduction and Fundamentals of Computational Swarm Intelligence.

Meet the Volume Editor

Dr. Jainath Yadav obtained an MTech and a Ph.D. from the Indian Institute of Technology Kharagpur. He joined the Central University of South Bihar in 2013. He is currently an Associate Professor at Central University of South Bihar, India. He served as the head of the Department of Computer Science at the Central University of South Bihar from 2020 to 2023. He has published several research papers in refereed journals and presented several papers at international conferences. He is the author of five books. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and a Ph.D. supervisor in his active research areas.

Contents

Preface

The edited volume is a collection of five reviewed chapters on recent advances in Association Rule Mining and Data Mining. Chapter 1 is an introductory chapter that provides a brief overview of association rule mining and data mining. Chapter 2 provides a systematic review of the use of data mining algorithms in chicken breeding. Chapter 3 discusses various data mining algorithms for predicting the body weight of non-descript indigenous goats.

Chapter 4 discusses automated data-driven and stochastic imputation methods to handle missing values in the dataset. Chapter 5 explores feature partitioning methods for data mining applications.

The book provides recent advances in association rule mining and data mining. We hope that readers will enjoy reading the book and also contribute to the data mining research community.

Dr. Jainath Yadav
Department of Computer Science,
Central University of South Bihar,
Gaya, India

Chapter 1

Introductory Chapter: Association Rule Mining and Data Mining

Jainath Yadav

1. Introduction

Data mining is the process of extracting meaningful patterns, trends, and insights from large datasets using various statistical, mathematical, and computational techniques [1]. It involves exploring and analyzing vast amounts of data to uncover hidden relationships, correlations, and anomalies that can be valuable for decision-making, prediction, and knowledge discovery. Data mining techniques include clustering, classification, association rule mining, regression analysis, and anomaly detection, among others [2]. By leveraging advanced algorithms and tools, data mining enables organizations to gain valuable insights into their data, leading to improved business strategies, enhanced customer experiences, and informed decision-making across various industries and domains [3]. Association rule mining serves as a cornerstone approach in the expansive realm of data mining, representing a fundamental approach to extracting valuable insights from complex datasets [4]. As an effective technique for revealing significant links and patterns that could otherwise go undetected, association rule mining stands out in today's data-driven world, when enterprises are deluged with enormous volumes of data. At its core, association rule mining enables analysts and researchers to identify associations between variables, revealing hidden connections and uncovering valuable knowledge buried within the data. This introductory segment sets the platform for a comprehensive exploration of association rule mining, laying the foundation for understanding its intricacies, applications, and implications in modern data analytics. By delving into the foundational concepts and principles that underpin association rule mining, readers gain a deeper appreciation for its significance and potential impact in driving informed decision-making processes across various domains and industries.

2. Foundations of association rule mining

The exploration of association rule mining commences with a comprehensive analysis of its fundamental concepts. Association rules, which indicate logical relationships between objects in a dataset, are the fundamental idea of association rule mining. In this segment, we delve into the nuances of association rules, exploring their definition, formulation, and significance in data mining endeavors [5]. Key metrics such as support, confidence, and lift emerge as essential indicators of the strength and reliability of association rules. Support quantifies the frequency of occurrence of a rule within the dataset, while confidence measures the degree of certainty that the presence of one item implies the presence of another. Lift, on the other hand, assesses

the degree of association between two items, accounting for the prevalence of both items in the dataset. Classic algorithms such as Apriori and FP-Growth take center stage in the exploration of association rule mining techniques. The Apriori algorithm laid the ground-work for efficient mining of frequent itemsets by utilizing a breadth-first search approach. FP-Growth is revolutionized association rule mining through its innovative use of a compact data structure known as the FP-tree, enabling efficient extraction of frequent patterns without the need for candidate generation [6].

3. Advances in algorithmic efficiency

Advancements in algorithmic efficiency have propelled association rule mining into new frontiers, enabling the analysis of massive datasets with unprecedented speed and scalability [7]. In this section, we embark on a journey to explore the latest developments in algorithm design, parallel and distributed computing techniques, and optimization strategies that have revolutionized association rule mining [8]. The quest for algorithmic efficiency is driven by the ever-growing volume, velocity, and variety of generated data. Traditional association rule mining algorithms, while effective, may struggle to cope with the computational demands imposed by large-scale datasets. As such, researchers have sought innovative solutions to enhance the speed and scalability of association rule mining algorithms, paving the way for more efficient data analysis processes. Parallel and distributed computing techniques emerge as key enablers of scalability, allowing association rule mining algorithms to leverage the computational power of multiple processors or distributed computing clusters. By parallelizing computationally intensive tasks, researchers can significantly reduce the time and resources required to perform association rule mining on large datasets, thereby unlocking new possibilities for data-driven insights and discoveries [9]. Optimization strategies play a pivotal role in enhancing the efficiency of association rule mining algorithms, enabling researchers to streamline data processing pipelines and minimize computational overhead. Techniques such as pruning, sampling, and data partitioning help researchers navigate the computational complexities inherent in association rule mining, ensuring that valuable insights can be extracted in a timely and resource-efficient manner [10].

4. Integration with other data mining techniques

Association rule mining thrives in synergy with other data mining techniques, amplifying the depth and breadth of insights garnered from data analysis endeavors. In this section, we delve into innovative approaches for integrating association rule mining with clustering, classification, and anomaly detection methods, showcasing the transformative impact of synergistic integration on data analysis processes. Clustering techniques such as k-means and hierarchical clustering are commonly used in conjunction with association rule mining to discover inherent patterns and structures inside datasets. By clustering similar data points together, researchers can uncover latent relationships and dependencies that may inform the discovery of association rules. Association rule mining, in turn, complements clustering by extracting actionable insights from the identified clusters, enabling researchers to uncover meaningful patterns and associations hidden within the data [11]. Classification algorithms, such as decision trees, support vector machines, and neural networks, play a crucial role in predictive modeling and pattern recognition tasks. By integrating

association rule mining with classification techniques, researchers can enhance the interpretability and accuracy of data predictive models, leveraging discovered association rules to inform the decision-making process. Association rules serve as valuable features or constraints in the classification process, providing insights into the relationships between input variables and target outcomes [12]. Anomaly detection techniques aim to identify outliers or deviations from normal patterns within datasets, serving as a critical component of fraud detection, intrusion detection, and quality control systems. By integrating association rule mining with anomaly detection methods, researchers can uncover anomalous patterns or behaviors that may signal potential fraud or irregularities within the data. Association rules serve as powerful indicators of anomalous behavior, enabling researchers to identify suspicious patterns and take appropriate remedial actions.

5. New perspectives and applications

The evolving landscape of association rule mining unveils a myriad of new perspectives and applications, heralding its relevance across an array of domains including retail, healthcare, finance, and telecommunications [13]. In this section, we embark on an exploration of emerging applications, ranging from market basket analysis and disease diagnosis to fraud detection and customer churn prediction. By illustrating the practical applications of association rule mining, readers gain insights into its transformative potential in driving informed decision-making and fostering innovation across diverse domains [14]. In the retail industry, association rule mining plays a pivotal role in market basket analysis, enabling retailers to uncover hidden patterns and associations within transactional data. By identifying frequently co-occurring items or product bundles, retailers can optimize product placement, tailor marketing strategies, and enhance the overall shopping experience for customers. Association rules serve as valuable insights into consumer behavior, guiding retailers in strategic decision-making processes and driving business growth. In the healthcare sector, association rule mining holds promise for improving disease diagnosis and treatment outcomes through the identification of meaningful patterns and associations within clinical datasets. By analyzing patient records, medical imaging data, and genomic data, researchers can uncover hidden relationships between symptoms, diagnoses, and treatment responses [15]. Association rules serve as valuable indicators of disease risk factors, treatment efficacy, and prognostic outcomes, empowering healthcare providers to deliver personalized and targeted interventions to patients. In the finance industry, association rule mining serves as a powerful tool for fraud detection and risk management, enabling financial institutions to identify suspicious patterns and behaviors within transactional data. By analyzing patterns of financial transactions, account activity, and user behavior, researchers can uncover anomalous patterns indicative of fraudulent activity or suspicious behavior. Association rules serve as valuable insights into fraudulent patterns, enabling financial institutions to mitigate risks, protect assets, and safeguard customer interests. In the telecommunications sector, association rule mining facilitates customer churn prediction and retention efforts through the identification of predictive patterns and behaviors within subscriber data. By analyzing call detail records, usage patterns, and customer demographics, researchers can uncover hidden relationships between customer characteristics and churn behavior. Association rules serve as valuable indicators of churn risk factors, enabling telecommunications providers to proactively engage with at-risk customers, tailor retention strategies, and enhance customer satisfaction and loyalty.

6. Challenges and future directions

Despite its remarkable advancements, association rule mining grapples with an array of challenges ranging from handling complex data types to ensuring privacy preservation in an era of evolving data landscapes. We go into great detail about these issues in this part and provide some future directions for association rule mining research and development. One of the primary limitations facing association rule mining is the handling of complex data types, including textual data, multimedia data, and spatiotemporal data. Traditional association rule mining algorithms are designed to operate on structured numerical or categorical data, posing challenges when confronted with unstructured or semi-structured data formats. Researchers are exploring innovative approaches for extending association rule mining techniques to handle diverse data types, leveraging techniques from natural language processing, computer vision, and geographic information systems. Privacy preservation emerges as another critical challenge in association rule mining, particularly in light of increasing concerns surrounding data privacy and security. As datasets grow larger and more diverse, the risk of inadvertently disclosing sensitive information or violating privacy regulations becomes more pronounced. Researchers are exploring techniques for privacy-preserving association rule mining, including differential privacy, secure multi-party computation, and homomorphic encryption, to ensure that sensitive information remains protected while still enabling meaningful analysis and insights to be gleaned from the data. Adapting to evolving data landscapes presents yet another challenge for association rule mining, as datasets continue to grow in size, complexity, and diversity. Traditional association rule mining algorithms may struggle to cope with the computational demands imposed by these evolving data landscapes, necessitating the development of more scalable and efficient techniques. Researchers are exploring novel approaches for parallel and distributed association rule mining, leveraging advances in cloud computing, distributed computing, and high-performance computing to enable efficient analysis of large-scale datasets. Interpretability and explainability emerge as key considerations in association rule mining, particularly in domains where transparency and accountability are paramount. Traditional association rule mining algorithms may produce large volumes of rules, making it challenging for stakeholders to interpret and understand the underlying patterns and associations. Researchers are exploring techniques for enhancing the interpretability and explainability of association rule mining results, including rule pruning, rule summarization, and visualization techniques, to facilitate meaningful interpretation and decision-making. Opportunities for research and innovation abound in association rule mining, as researchers continue to explore new methodologies, techniques, and applications for extracting actionable insights from data. By addressing the challenges posed by complex data types, privacy preservation, evolving data landscapes, and interpretability, researchers can unlock new frontiers in association rule mining and pave the way for its continued relevance and impact in modern data analytics.

7. Conclusion

Association rule mining stands as a linchpin within the data mining toolkit, enabling the discovery of valuable insights and patterns in large datasets. By unraveling its foundational principles, recent advances, applications, and challenges,

practitioners and researchers are poised to harness the power of association rule mining to drive informed decision-making and unlock the potential of data-driven innovation. As the data landscape continues to evolve, association rule mining remains a beacon of opportunity, empowering stakeholders to extract actionable insights and drive transformative change across diverse domains and industries. Through a concerted effort to address the challenges and opportunities presented by association rule mining, researchers can unlock new frontiers in data analytics and pave the way for a future where data-driven insights drive innovation, growth, and progress.

Author details

Jainath Yadav
Department of Computer Science, Central University of South Bihar, Bihar, India

*Address all correspondence to: jaibhu38@gmail.com

IntechOpen

References

[1] Adriaans P. Data mining. Pearson Education India; 1996

[2] Chen M-S, Han J, Yu PS. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering. 1996;**8**(6):866-883

[3] Gul S, Bano S, Shah T. Exploring data mining: Facets and emerging trends. Digital Library Perspectives. 2021;**37**(4):429-448

[4] Telikani A, Gandomi AH, Shahbahrami A. A survey of evolutionary computation for association rule mining. Information Sciences. 2020;**524**:318-352

[5] Xu S. Deep mining method for high-dimensional big data based on association rule. International Journal of Internet Protocol Technology. 2021;**14**(3):147-154

[6] Yuan X. An improved apriori algorithm for mining association rules. In: AIP Conference Proceedings. Vol. 1820. United States: AIP Publishing; 2017

[7] Al-Maolegi M, Arkok B. An improved apriori algorithm for association rules. International Journal on Natural Language Computing. 2014;**3**:21-29

[8] Zhao Z, Jian Z, Gaba GS, Alroobaea R, Masud M, Rubaiee S. An improved association rule mining algorithm for large data. Journal of Intelligent Systems. 2021;**30**(1):750-762

[9] Liu L, Wen J, Zheng Z, Su H. An improved approach for mining association rules in parallel using spark streaming. International Journal of Circuit Theory and Applications. 2021;**49**(4):1028-1039

[10] Sharmila S, Vijayarani S. Association rule mining using fuzzy logic and whale optimization algorithm. Soft Computing. 2021;**25**(2):1431-1446

[11] Dol SM, Jawandhiya PM. Classification technique and its combination with clustering and association rule mining in educational data mining—A survey. Engineering Applications of Artificial Intelligence. 2023;**122**:106071

[12] Veerappa M, Anneken M, Burkart N, Huber MF. Explaining cnn classifier using association rule mining methods on time-series. In: Explainable Deep Learning AI. London: Elsevier; 2023. pp. 173-189

[13] Tsui K-L, Chen V, Jiang W, Yang F, Kan C. Data mining methods and applications. In: Springer Handbook of Engineering Statistics. London: Springer; 2023. pp. 797-816

[14] Diaz-Garcia JA, Ruiz MD, Martin-Bautista MJ. A survey on the use of association rules mining techniques in textual social media. Artificial Intelligence Review. 2023;**56**(2): 1175-1200

[15] Agapito G, Guzzi PH, Cannataro M. Parallel and distributed association rule mining in life science: A novel parallel algorithm to mine genomics data. Information Sciences. 2021;**575**:747-761

Chapter 2

Use of Data Mining Algorithms in Chicken Breeding: A Systematic Review

Thobela Louis Tyasi, Madumetja Cyril Mathapo, Kwena Mokoena, Victoria Rankotsane Hlokoe and Kagisho Madikadike Molabe

Abstract

Data mining algorithms have been performed to reveal the factors that can be used to enhance live body weight and egg weight during chicken breeding. This work was conducted to systematically review the published articles on the use of data mining algorithms in chicken breeding. ScienceDirect, Web of Science, PubMed, Google Scholar and were used for searching articles. Using the combination of chicken or chicken breeding, data mining algorithm or decision tree, body weight and egg weight as keywords. The results indicated that 8 articles were included from 120 articles were found from searching. The 8 included articles were published from 2016 to 2021 and most of them were originated from South Africa (n = 3) followed by Turkey (n = 2) with. CHAID as the most used data mining algorithm (n = 5) followed by CART (n = 4). Out of 8 included articles, 6 of them used coefficient of determination (R^2) as the selection criteria and CART was found as the best model followed by the CHAID model. It is concluded that CART followed by CHAID data mining algorithms are the recommended models that might be used for improving egg production and growth performance of chickens.

Keywords: body weight, coefficient of determination, chicken breeding, data mining algorithm, egg weight

1. Introduction

Chicken breeding focuses on improving different animal productions including the growth performance, carcass characteristics and egg production. Different studies had been conducted trying to improve growth performance [1–3] and egg production [1, 4–7] using different data mining algorithms. Data mining algorithms are nonparametric methods superior and simpler in statistically calculating complex data sets [3]. Moreover, Gevrekçi and Takma [5] reported that they are computer-based procedures to detect evidence from data removing multicollinearity and can run large data. The common data mining algorithms that are performed for estimation of chicken

live body weight are classification and regression tree (CART) and artificial neural network (ANN) in Sasso breed [1], and exhaustive chi-square automatic interaction detector (exhaustive CHAID) and chi-square automatic interaction detector (CHAID) in Hy-line Silver Brown and Potchefstroom Koekoek chicken breed [3] and multivariate adaptive regression splines (MARS) in Hy-line Silver Brown chicken [2]. chi-square automatic interaction detector (CHAID) in Hy-line Silver Brown and Boschveld layers [8] and in White layer hybrids chicken [4], chi-square automatic interaction detector (CHAID) and classification and regression tree (CART) in Indigenous chicken of Zambia [7], k-nearest neighbor (KNN), linear discriminant analysis (LDA) and Support vector machine (SVM) in Beijing You Chicken and Dwarf Beijing You Chicken [6], and chi-square automatic interaction detector (CHAID) and ridge regression (RR) in White layer hybrids [4]. Chi-square automatic interaction detector (CHAID, and Classification and regression tree (CART) are commonly performed algorithm methods to improve egg production [5].

Based on authors knowledge, there is no systematic review on the use of data mining algorithms in chicken breeding. To close the identified knowledge gap, the objective of this work was to perform the systematic approach to review the information on the use of data mining algorithms in chicken breeding. This book chapter will help the chicken breeders and researchers to identify the potential data mining algorithms that might be used for estimation of live body weight and egg weight.

2. Methods and materials

2.1 Eligibility criteria

The Population, Exposure and Outcomes (PEO) as components were identified as outlined by Saltikov [9]. The "Chicken" was defined as population of the study, while the "Data mining algorithm or decision tree" as intervention, "Eggs weight" and "Body weight" as outcome. A preliminary search of the PEO component on Google Scholar, Web of Science, PubMed and ScienceDirect was performed before deciding to conduct the study.

2.2 Search strategy

A scientific publication search was performed independently by two investigators (Kwena Mokoena and Thobela Louis Tyasi) in databases up to 10th November 2023, using Google Scholar, Web of Science, PubMed and ScienceDirect. The search was performed using the combination of keywords as follows: 'Chicken" or "Chicken breeding", "Data mining algorithm" or "Decision tree"," Body weight", and "Egg weight".

2.3 Inclusion criteria

Searched articles were selected for eligibility according to several standard and considered for inclusion if they met the following criteria:

- Chicken
- Data mining algorithm or decision tree

- Egg weight
- Live body weight

2.4 Exclusion criteria

The criteria of excluding searched articles contained the following:

- Records irrelevant to data mining algorithm, egg weight, carcass weight and body weight
- Studies published as abstract without full text
- Records duplicated
- Studies not on chickens
- Articles with no available original data in the publication and failure to contact the authors

2.5 Data extraction

The data for the current study was extracted independently by Kwena Mokoena and Thobela Louis Tyasi, and an agreement was made involving all sections. The information obtained from each article consisted of the following:

- First author
- Year of publication
- Number of eggs weight
- Chicken breed
- Data mining algorithm or decision tree
- Dependent variables (egg weight and live body weight).

2.6 Ethical considerations

When performing this work all authors considered plagiarism, fabrication, and data falsification.

3. Results

3.1 Searched results

One hundred and twenty (n = 120) articles were retrieved from a publication search, were twenty-five (n = 25) of which were duplicated were removed. As a result,

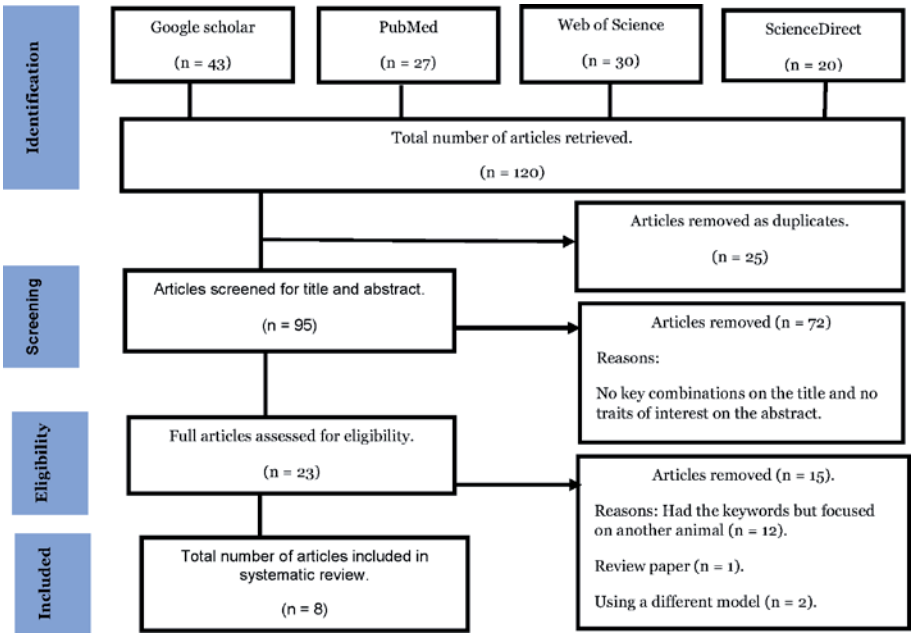

Figure 1.
Flowchart of identification and selection of studies for systematic review.

ninety-five (n = 95) articles were considered for title and abstract screening, which resulted in seventy-two (n = 72) articles eliminated after title and abstract review. Twenty-three (n = 23) articles were considered for full text-review, a total of fifteen (n = 15) articles were eliminated after a full text- review, the reasons are stated in **Figure 1**. A total of eight (n = 8) articles qualified for the inclusion in the study.

3.2 Characterization of included articles

Table 1 shows eight articles that met the inclusion procedure. The results indicated that [2–5, 8] used commercial chicken breeds and their eggs. The study that used large sample size of chickens was [8], while the study that used large sample size of chicken eggs was of [4]. The results showed that the most dominant chicken breed was Hy-line silver, Brown layer [2, 3, 5, 8].

3.3 Publication by year

Figure 2 indicates the year of publication of included articles. The findings indicated that year 2017 [1, 3, 7, 8] had the highest numbers of articles published (n = 2). The year 2016 [4] and 2018 [5] showed the least number of articles.

3.4 Publication by county

The origin of the included articles is presented in **Figure 3**. The results indicated that from the eight articles included in the review, South Africa [2, 3, 8] had the maximum number of articles (n = 3) and followed by Turkey [5, 8] with two articles. The results also indicated that Nigeria [1], Zambia [7] and China [6].

Authors	Years	Country	Breeds	Data mining algorithm
Gevrekçi and Takma	2018	Turkey	—	Classification and regression tree (CART), and chi-square automatic interaction detector (CHAID)
Dong et al.	2021	China	Beijing You Chicken and Dwarf Beijing You Chicken	k-nearest neighbor (KNN), linear discriminant analysis (LDA) and Support vector machine (SVM)
Liswaniso et al	2020	Zambia	Indigenous chicken of Zambia	Classification and regression tree (CART), and chi-square automatic interaction detector (CHAID)
Okoro et al.	2017	South Africa	Hy-line Silver Brown and Boschveld layers	chi-square automatic interaction detector (CHAID)
Orhan et al	2016	Turkey	White layer hybrids	chi-square automatic interaction detector (CHAID) and ridge regression (RR).
Tyasi et al.	2020	South Africa	Hy-line Silver Brown	Multivariate adaptive regression splines (MARS).
Tyasi et al	2021	South Africa	Hy-line Silver Brown and Potchefstroom Koekoek	Classification and regression tree (CART), Chi-square automatic interaction detector (CHAID) and exhaustive chi-square automatic interaction detector (exhaustive CHAID).
Yakubu and Madaki	2017	Nigeria	Sasso breed	Artificial neural network (ANN), Classification and regression tree (CART)

Table 1.
Characteristics of included studies.

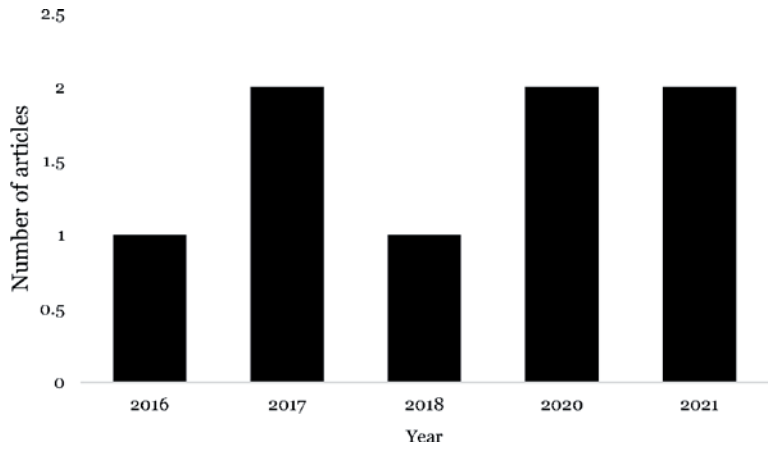

Figure 2.
Publication by year.

3.5 Publications by data mining algorithms

Figure 4 displays the number of published articles by data mining algorithms. The findings showed that CHAID was the more commonly used data mining algorithm

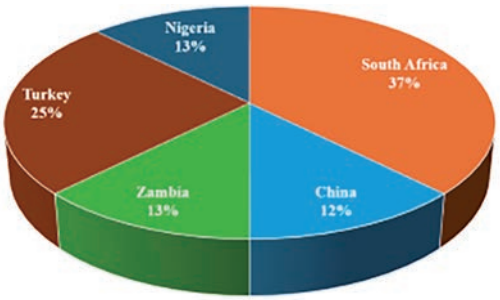

Figure 3.
Publications by a country.

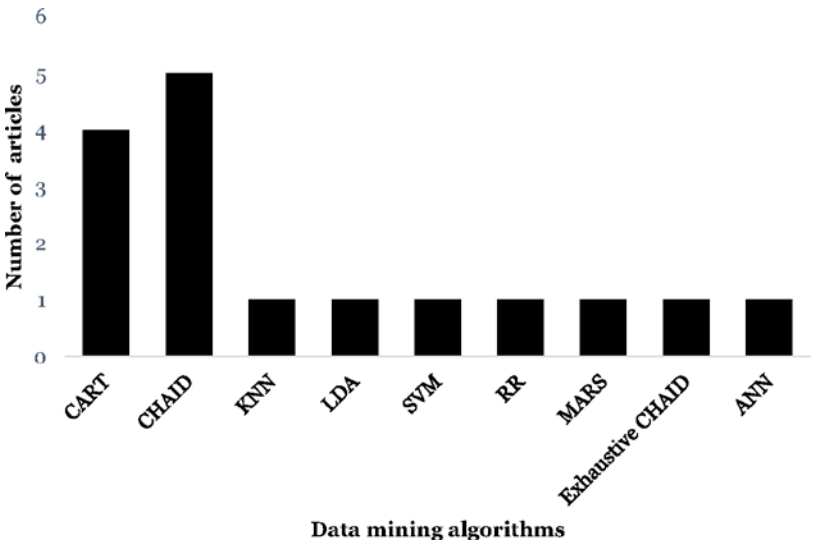

Figure 4.
Publications by data mining algorithms.

(n = 5), followed by CART (n = 4). The results also indicated that KNN, LDA, SVM, RR, MARS, Exhaustive CHAID and ANN were the data mining algorithms to be used (n = 1).

3.6 Predictive performance of data mining algorithms

Table 2 displays the predictive performance of different data mining algorithms used in the included articles for this study. From eight included articles only seven articles reported goodness of fit. Out of seven studies, most of them (six) used coefficient of determination (R^2) as the selection criteria. However, only one study used CV, RAE, MAD and RMSE [5]. From the six studies that used R^2 as the selection criteria, it was found that CART was the best model, then followed by the CHAID model. The study of Gevrekçi and Takma [5] indicated that CHAID was the best data mining algorithm model.

Author and year	Dependent variable	Goodness of fit criteria	Models								
			CART	CHAID	Exhaustive CHAID	RR	MARS	ANN	KNN	LDA	SVM
Gevrekçi and Takma, 2018	Egg production	CV%	10.57	9.32	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		RAE	0.0024	0.0021	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		MAD	8.85	7.56	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		RMSE	11.25	9.93	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dong et al. 2021	Egg discrimination (fatty acid)	R^2	N/A	N/A	N/A	N/A	N/A	N/A	91.7%	83.3%	91.7%
Dong et al. 2021	egg discrimination (flavor characteristics)	R^2	N/A	N/A	N/A	N/A	N/A	N/A	50%	N/A	16.7%
Liswaniso et al. 2020	Egg weight	R^2	59.3%	82.3%	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Orhan et al. 2016	Egg weight	R^2	N/A	99.98%	N/A	93.15%	N/A	N/A	N/A	N/A	N/A
Tyasi et al. 2020	Body weight	R^2	N/A	N/A	N/A	N/A	100%	N/A	N/A	N/A	N/A
Yakubu and Madaki, 2017	Body weight (deep litter)	R^2	93.4%	N/A	N/A	N/A	N/A	87%	N/A	N/A	N/A
Yakubu and Madaki, 2017	Body weight (battery cage)	R^2	93.4%	N/A	N/A	N/A	N/A	99%	N/A	N/A	N/A
Tyasi et al. 2021	Body weight	R^2	83.2%	65.9%	64.1%	N/A	N/A	N/A	N/A	N/A	N/A
Okoro et al. 2017	Egg size and performance	R^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

CART: Classification and regression tree; CHAID: Chi-square automatic interaction detector; RR: Ridge regression; MARS: Multivariate adaptive regression splines; ANN: Artificial neural network; KNN: K-nearest neighbor; LDA: linear discriminant analysis; SVM: Support vector machine; R^2: Coefficient of determination; CV: Coefficient of variation; RAE: Relative approximate error; MAD: Mean absolute deviation; RMSE: Root mean square error; N/A: Not applicable.

Table 2.
Goodness of fit criteria for data mining algorithms in prediction of dependent variables.

4. Discussion

This review was conducted to discover the suitable data mining algorithm model that might be used in chicken breeding from 8 included articles. The findings showed that CHAID was the most used data mining algorithm (5/8) out of the eight articles included in the review, followed by CART (4/8). However, the predictive performance results indicated six articles from included studies used the coefficient of determination (R^2) as the selection criteria. This shows that R^2 is the reliable goodness of fit criteria for selecting the best model. However, the CV, RAE, MAD and RMSE were used by only one article [5]. From the six articles that used R^2 as the selection criteria, it was found that CART was the best model, then followed by the CHAID model. CART algorithm is a kind of machine learning technique performed to assemble a decision tree [2]. The study of Gevrekçi and Takma [5] indicated that CHAID was the best data mining algorithm model. Liswaniso et al. [7] used the different data mining algorithm models to determine egg weight as dependent variable from egg characteristics and found that CHAID is the best model. Similarly, Tyasi et al. [3] found that CHAID model is the best in predicting the body weights of Hy-line Silver Brown commercial layers and Potchefstroom Koekoek indigenous chickens raised in South Africa. To the best of authors' knowledge, this is the first review in a systematic approach reporting the use of data mining algorithms in chicken breeding. Hence, there is no comparison of our findings in this systematic review. The implication of this work is that the CART method might be used for prediction of live body weight and egg weight for growth performance and egg production improvement in different countries. The strength of the review is that there is no similar study had been done. The contribution of this systematic review is that out of all the commonly used statistical technique for prediction of egg weight and live body weight in chickens, CART is the best model that can be used during chicken breeding. However, more studies need to be done to confirm and add to the results of the study.

5. Conclusion

The current systematic review was conducted to discover the best data mining algorithm that might be used by chicken breeders to identify the factors for improving live body weight and egg weight. Included articles identified factors that can be used during breeding as selection criteria to improve egg production and growth performance. This systematic review showed that included articles used different data mining algorithms including classification and regression tree (CART), artificial neural network (ANN), chi-square automatic interaction detector (CHAID), exhaustive chi-square automatic interaction detector (exhaustive CHAID), multivariate adaptive regression splines (MARS), k-nearest neighbor (KNN), linear discriminant analysis (LDA), support vector machine (SVM) and ridge regression (RR) for chicken breeding. Included articles used goodness of fit criteria such as coefficient of determination and root mean square error to select the best data mining algorithm. This systematic review concludes that CART was the best data mining algorithm model to be used in chicken breeding, followed by CHAID. Furthermore, the researchers should involve the CART and CHAID methods in chicken breeding for prediction of egg weight and live body weight.

Conflict of interest

The authors declare no competing of interests.

Declarations

We declare that this is our work.

Author details

Thobela Louis Tyasi*, Madumetja Cyril Mathapo, Kwena Mokoena, Victoria Rankotsane Hlokoe and Kagisho Madikadike Molabe
Department of Agricultural Economics and Animal Production, University of Limpopo, Sovenga, Limpopo, South Africa

*Address all correspondence to: louis.tyasi@ul.ac.za

References

[1] Yakubu A, Madaki J. Modelling growth of dual-purpose Sasso hens in the tropics using different algorithms. Journal Genetic Biology. 2017;**1**(1):1-9

[2] Tyasi TL, Makgowo KM, Mokoena K, Rashijane LT, Mathapo MC, Danguru LW, et al. Multivariate adaptive regression splines data mining algorithm for Presiction of body weight of Hy-line silver Brown commercial layer chicken breed. Advances in Animal and Veterinary Sciences. 2020;**8**(8):794-799. DOI: 10.17582/journal.aavs/2020/8.8.794.799

[3] Tyasi TL, Eyduran E, Celik S. Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa. Tropical Animal Health and Production. 2021;**53**(7):1-8. DOI: 10.1007/s11250-020-02443-y

[4] Orhan H, Eyduran E, Tatliye A, Saygici H. Prediction of egg weight from egg quality characteristics via ridge regression and regression tree methods. 2016;**45**(7):380-385

[5] Gevrekci Y, Takma C. A comparative study for egg production in layers by decision tree analysis. Pakistan Journal of Zoology. 2018;**50**(2):437-444. DOI: 10.17582/journal.pjz/2018.50.2.437.444

[6] Dong X, Gao L, Zhang H, Wang J, Qiu K, Qi G, et al. Discriminating eggs from two local breeds based on fatty acid profile and flavor characteristics combined with classification algorithms. Food Science of Animal Resources. 2021;**41**(6):936-949. DOI: 10.5851/kosfa.2021.e47

[7] Liswaniso S, Qin N, Tyasi TL, Chimbaka IM, Sun X, Xu R. Use of data mining algorithms Chaid and CART in predicting egg weight from egg quality traits of indigenous free-range chickens in Zambia. Advances in Animal and Veterinary Sciences. 2020;**9**(2):215-220. DOI: 10.17582/journal.aavs/2021/9.2.215.220

[8] Okoro VMO, Ravhuhali KE, Mapholi TH, Mbajiorgu EF, Mbajiorgu AC. Comparison of commercial and locally developed layers performance and egg size prediction using regression tree methods. The Journal of Applied Poultry Research. 2017;**26**:477-484

[9] Bettany-Saltikov J. Learning how to undertake a systematic review: Part 2. Nursing Standard. 2010;**24**:47-56

Chapter 3

Comparison of MARS, CART, and Linear Regression Models for Prediction of Body Weight of Non-descript Indigenous Goats in Lepelle-Nkumbi Local Municipality, South Africa

Madumetja Cyril Mathapo, Thobela Louis Tyasi, Kwena Mokoena, Rankotsane Victoria Hlokoe and Madikadike Kagisho Molabe

Abstract

In Lepelle-Nkumbi Local Municipality of South Africa, 200 none-descript indigenous goats ranging in age from one to five years were the subjects of a study that compared the live body weight predictions made by stepwise linear regression, Classification Regression Tree (CART), and Multivariate Adaptive Splines (MARS) models. Several bodily measurements, such as canonical circumference (CC), sternum height (SH), body length (BL), ear length (EL), head length (HL), head width (HW), rump length (RL), rump height (RH), and rump width (RW). The evaluation criteria included the root mean square error (RMSE), coefficient of determination (R^2), to decide which model was the best. According to the results, CART outperformed the others, obtaining the lowest RMSE (3.65) and the greatest R^2 (0.80). The stepwise regression model outperformed data mining algorithms in male goats. According to the study, CART is a useful statistical technique for defining requirements for producing indigenous goats that are not very special. In addition, when predicting live body weight from body measuring features, the stepwise regression model should be considered.

Keywords: MARS, CART, stepwise regression model, goodness of fit, indigenous goats

1. Introduction

Livestock body weight can assist farmers in accurately administering medication (drug dosage), providing optimum feeding, estimating market prices, and deciding on

appropriate breeding strategy to be implement [1]. In remote areas where there's lack of weighing scales, morphological traits serve as the simple and cheap method for estimation of body weight [2]. Numerous investigations have been carried out on predicting body weight based on morphological traits using linear regression models [3, 4]. However, it was indicated that this linear regression models fail to accurately estimate live body weight of the animals since they are not able to detect and overcome multi-collinearity problems that occurs between independent variables [5]. The focus on enhancing developmental breeding strategies has increased the popularity of data mining algorithms such as Classification and Regression Tree (CART) and Multivariate Adaptive Splines (MARS) [6], and their application for estimation of body weight has been applied efficaciously on different livestock [7, 8]. CART and MARS are non-parametric methods and are employed for statistical analysis for ordinal, nominal and continuous variables to discover the effect of explanatory variables on categorical response variables [9]. They are statistical techniques which study the mathematical relation between one or more explanatory and response variables [10]. The utilization of CART and MARS methodologies in predicting animal body weight is crucial, considering various elements including age, breed, sex, and environment, all of which can influence body weight [11]. To the best of the authors knowledge, there is a dearth of information comparing stepwise linear regression, MARS, and CART to estimate the live body weight of non-descript indigenous goats in South Africa. Assessing and contrasting MARS, CART, and stepwise linear regression models for predicting live body weight of non-descript indigenous goats in South Africa was the aim of the study. The goal of the study is to determine which model is best suited for measuring the live body weight of non-descript indigenous goats in South Africa, as well as what factors are most important in raising live body weight.

2. Methods and materials

2.1 Study area

The present study took place in four communities in Lepelle-Nkumpi Local Municipality, which is situated in the Capricorn District of Limpopo Province, South Africa, were the sites of the current study. Lepelle-Nkumpi's geographic coordinates are roughly 24° 14′60.00″S latitude and 29° 39′59.99″E longitude. The average annual temperature of the area is 20°C, with summer temperature typically reaching 23°C and winter temperature at 20°C. The area receives between 453 and 474 mm of rain annually [12].

2.2 Experimental animals, management, and design

Non-descript indigenous goats of age between 1 to 5 years, sourced from villages within Lepelle-Nkumpi Local Municipality were used. An extensive farming system was employed, and non-pregnant, healthy goats were randomly selected for inclusion in the study. Cross sectional design was employed in the study.

2.3 Sampling method and the size of the sample

A multi-stage sampling method was utilized, with the deliberate selection of Lepelle-Nkumpi Local Municipality due to its highest population of non-descript indigenous

goats, as reported by the Department of Agriculture, Land Reform Development in Limpopo. Four villages, namely: Lenting, Morotse, Seleteng and Lesetsi were purposively selected due to the extension officer working hand in hand with the farmers from those villages. Five farmers with at least 10 non-descript indigenous goats in their herd from each village were randomly selected. Therefore, a total of 200 goats of different sexes were used for linear body measurements traits in the study.

2.4 Data collection

A weighing machine calibrated in kilograms (kg) was used to determine the live body weight (BW), and tailor tape calibrated in centimeters (cm) was used to measure the body weight. Using Yakubu's method [7], the following linear body measurements were obtained: muzzle diameter (MD), head length (HL), withers height (WH), sternum height (SH), body length (BL), head width (HW), heart girth (HG), ear length (EL), rump length (RL), rump height (RH), rump width (RW), and canonical circumference (CC).

2.5 Classification and regression tree (CART) and multivariate adaptive regression splines (MARS)

The binary decision tree structure known as CART is produced by recursively splitting a parent node, which at first holds the whole dataset, into two child nodes. According to Tyasi et al. [13], several comparable nodes are produced from a learning dataset using cross-validation training and test sets to minimize error variation. Producing a terminal node that improves node differentiation is the aim of this methodology [14]. The goal of CART is to create a model that can be easily understood for both ordinal and nominal scale event prediction [15]. In contrast, MARS is referred to as a non-parametric regression technique. The MARS algorithm was applied in accordance with Sengul et al. [6], explanations.

2.6 Stepwise linear regression

The following procedure was followed to estimate body weight via linear body measures using a stepwise regression model:

$$BW = a + b1x1 + b2x2 + \ ... \ + bnxn \tag{1}$$

Where:
a = regression intercept.
b's = regression coefficient of linear body measurements.
X's = independent (WH, RH, HG, SH, BL, HL, HW, EL, MD, RL, RW, CC).
e = random error term.
BW = dependent (BW).

2.7 Goodness of fit criteria

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{2}$$

$$r = \frac{cov\left(y_i y_{ip}\right)}{S_{yi} S_{Yip}} \tag{3}$$

$$Adj.R^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{4}$$

$$SD_{ratio} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}}{y_i}\right|.100 \tag{6}$$

$$AIC = nln\left[\sum_{i=1}^{n}\frac{\left(y_i - \hat{y}_i\right)^2}{n}\right] + 2k \tag{7}$$

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_I - \hat{Y_I}\right)}{\sum_{i=1}^{n} Y_I^2}} \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

$$CV(\%) = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_I - \bar{\varepsilon}\right)^2}{\overline{Y}}} \times 100 \tag{10}$$

$$PI = \frac{rRMSE}{1 + r} \tag{11}$$

$$MAD = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\frac{Yi - \hat{Y}i}{Yi}\right| \times 100 \tag{12}$$

2.8 Data analysis

The development Version 29.0 of the Statistical Package for the Social Sciences was used in the creation of the stepwise regression model. To build and assess the MARS and CART models prediction abilities, Rstudio was used.

3. Results

3.1 MARS model

The MARS model developed to predict BW had 4 basic functions (**Table 1**). The model commenced with an intercept term, featuring a regression coefficient of 29.45. The subsequent term, being the first basic function, indicated that heart girth (HG) is greater than 65 cm, with a regression coefficient of 0.788.

Basic functions (BF)		Coefficient
Intercept	Intercept	29.45
BF1	h(HG-65)	0.788
BF2	h(MD -19)	−0.734
BF3	h(15-RL)	−2.38
BF4	h(RH-57)	0.41

HG: Heart girth; MD: Muscle diameter; RL: Rump length; RH: Rump height.

Table 1.
MARS model.

3.2 CART model

The diagram created using the regression and classification tree model is displayed in **Figure 1**. The initial node in the diagram displayed an average body weight (BW) of 39 kg, representing the entirely of the (100%). This root node was subsequently split into two subgroups. The primary influential variable on BW at the first level was heart girth (HG). Following that, the second level variables were RH and HG, the third level variables were BL, WH, and age (three years), the fourth level variable was RH, the fifth level variable was HG, the sixth level variable were WH and MD, and the seventh level variable was SH. The division of the root node was based on HG, specifically with value of 75 cm. from the first tree depth where HG is <75 cm the average BW was found to be 33 kg representing 50% of the animals while HG > 75 cm had the average BW of 44 kg representing 50% of the animals. At the second tree depth where RH was <58 cm the average BW was 27 kg representing 9% of the animals and where the RH > 58 cm, the average BW was 35 kg representing 41% of the animals. While when HG < 81 cm the average BW was 41 kg representing 26% and when HG was >81 cm the average BW was 48 kg representing 24% of the animals. At the third tree depth where BL was <60 cm the average BW was 22 kg representing 3% of the animals and when BL was >60 cm the average BW was 29 kg representing 6%, and when WH ≥ 56 cm the average BW was

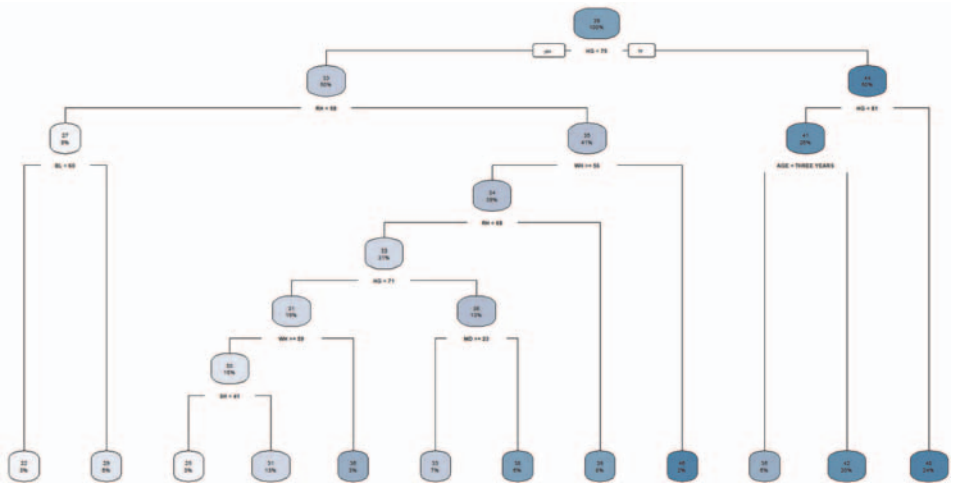

Figure 1.
Showing CART diagram.

46 kg with 2% representing the animals and when WH < 56 cm the average BW was 34 kg, representing 39% of the animals. Where the age was equal to three years the average BW was 35 kg representing 6% of the animal and 42 kg representing 20% of the animals. At the fourth tree depth where RH was <68 cm the average BW was 33 kg representing 31% of the animals and when RH > 68 cm the average BW was 39 kg representing 8% of the animals. At the fifth tree depth were HG < 71 cm the average BW was 31 kg representing 19% of the goats while when HG > 71 cm the averages BW was 36 kg with 13% representing the animals. At the sixth tree depth where WH was ≥59 cm the average BW was 36 kg representing 3% of the animals while when WH < 59 cm the average BW 30 kg with representing 15% of the animal. While where MD ≥ 23 cm the average BW was 38 kg representing 6% of the animals and when MD < 23 cm the average BW was 33 kg representing 7% of the animals. At the seventh tree depth where SH < 41 cm the average BW was 25 kg representing 3% of the animals and when SH was >41 cm, the average BW was 31 kg representing 13% of the animals.

3.3 The predictive accuracy of MARS and CART

The goodness of fit indicated how well MARS and CART performed on both the training and test datasets (**Table 2**). The findings indicated that CART was the best model in terms of training data set. The findings showed that CART model had the lowest RMSE (3.65), RRMSE (9.43), SDR (0.45), CV (9.46), RAE (0.01), MAPE (7.93), AIC (370.44) and highest R^2 (0.80), AdjR^2 (0.80) and PC (0.89).

3.4 Regression analysis

The stepwise regression model for female goats is shown in **Table 3**. The stepwise regression model includes features like RW, RH, SH, HW, and CC. the outcomes

Criterion	CART		MARS		Decision
	Train	Test	Train	Test	
RMSE	3.65	6.28	4.36	5.49	Small is fantastic
RRMSE	9.43	16.01	11.24	13.99	Small is fantastic
SDR	0.45	0.82	0.53	0.71	Small is fantastic
CV	9.46	16.13	11.28	14.05	Small is fantastic
PC	0.89	0.65	0.85	0.73	High is fantastic
RAE	0.01	0.03	0.01	0.02	Small is fantastic
MAPE	7.93	11.03	8.46	10.01	Small is fantastic
R^2	0.80	0.34	0.72	0.49	High is fantastic
AdjR^2	0.80	0.34	0.71	0.44	High is fantastic
AIC	370.44	216.80	430	210.86	Small is fantastic

RRMSE: Relative root means square error; RMSE: Root mean square error; CV: Coefficient of variation; AdjR^2: Adjusted coefficient of determination; SDR: Standard deviation ratio; PC: Pearson correlation; MAPE: Mean approximate error; R^2: Coefficient of determination; AIC: Akaike information criterion; RAE: Relative approximate error.

Table 2.
The predictive accuracy of MARS and CART models.

Model	α	β1	β2	β3	β4	β5	β6	R^2	AdjR^2	AIC	BIC	RMSE
RW	7.67	1.97						0.36	0.36	713.25	716.50	6.41
RW + RH	−15.21	1.38	0.49					0.47	0.47	678.37	684.89	5.84
RW + RH + HG	−22.98	1.19	0.41	0.22				0.53	0.53	657.44	667.22	5.51
RW + RH + HG + SH	−32.97	0.90	0.27	0.29	0.40			0.59	0.58	633.97	647.00	5.17
RW + RH + HG + SH + HW	−36.26	0.68	0.20	0.30	0.33	1.18		0.62	0.61	620.47	636.76	4.98
RW + RH + HG + SH + HW + CC	−37.82	0.58	0.15	0.28	0.30	1.04	1.24	0.64	0.62	615.79	635.34	4.91

RH: Rump height; RW: Rump width, HW: Head width; HG: Heart girth; SH: Sternum height; CC: Canon circumference; α: constant; AdjR^2: Adjusted coefficient of determination β1: variable; AIC: Akaike Information Criteria; R^2: Coefficient of determination; BIC: Bayesian Information Criteria; RMSE: Root means square error.

Table 3.
Using stepwise regression to estimate the body weight of female goats based on selected linear body measurements.

Model	α	β2	β2	R^2	$AdjR^2$	AIC	BIC	RMSE
HG	−46.03	1.12		0.92	0.91	20.63	23.88	2.84
HG + CC	−52.96	0.93	2.32	0.97	0.97	12.03	12.64	1.79

R^2: Coefficient of determination; HG: Heart girth; CC: Canon circumference; RMSE: Root means square error; α: constant; β1: variable; AIC: Akaike Information Criteria; BIC: Bayesian Information Criteria; $AdjR^2$: Adjusted coefficient of determination.

Table 4.
Predicting the body weight of male goats using stepwise regression based on specific linear body measurements.

demonstrated that RW was the first characteristic to be incorporated into the model, explaining minimal variation in BW with a high AIC (713.25), BIC (716.0), and RMSE (6.41) and a lower R^2 (36%). The findings showed that AIC, BIC, and RMSE decreased while R^2 and $AdjR^2$ increased with the addition of more selected traits to the model. This pattern demonstrated the models increased predictive power for BW.

Table 4 displays a stepwise regression model for male goats. Linear body measurements such as HG and CC were incorporated into the model. The findings indicated HG as the first trait to be included in the model and explained a higher variation in BW of male goats with R^2 (92%), $AdjR^2$ (0.91), AIC (20.63), BIC (23.88) and RMSE (2.84). The results further showed a higher variation, explained in the model when CC was included, with R^2 (97%), $AdjR^2$ (0.97), AIC (12.03), BIC (12.64) and RMSE (1.79). These results indicated that as traits added on HG the accuracy of the model for prediction of BW also increases.

4. Discussion

Different statistical approaches can be utilized to elucidate the variance between distinct characteristics and body weight [16]. Linear body measurements performs a crucial role in estimating live body weight, achieving an accuracy level of up to 90% compared to the actual body weight [13]. CART and MARS, along with a stepwise regression model, were initially employed to determine the impact of linear body measurement traits on the body weight of non-descript indigenous goats. The MARS model identified heart girth exceeding 65 cm as the most reliable estimator of live body weight for these goats. Similarly, CART model also highlighted heart girth as a primary estimator for live body weight for these goats. In does, stepwise regression models suggested that rump width could be employed as a sole trait for predicting body weight, while in the male goats, heart girth emerged as the most effective single predictor. The study's findings contrast with those of Faraz et al. [17], where body length was identified as the top predictor for Thalli sheep. This discrepancy may be attributed to variations in environmental conditions and species. However, the current study aligns with the conclusions of Berhe [18], Temoso et al. [19], and Odadi [20], where heart girth was identified as the most reliable estimator of live body weight in both goats and sheep. Predictive performance findings indicated that CART model had the best predictive performance compared to MARS and stepwise regression models. However, it was outperformed by stepwise regression models in male goats. The findings of Celik [21] and Tirink et al. [16] were similar with the findings of the

current study where CART performed better than MARS. However, Celik [22] and Faraz et al. [17], reported different findings from current study where it was found that CART model performed poor as compared to MARS in Pakistan goats and Thaili sheep, respectively.

5. Conclusion

Based on the results of the MARS, CART, and stepwise regression models, the greatest predictive live body weight, according to the current study, is heart circumference. When it came to forecasting the body weight of ordinary native goats of South Africa, the CART model outperformed the other models, according to the predictive performance. On the other hand, stepwise regression model outperformed data mining methods in male goats. These results point to the CART model's importance as a useful technique for setting guidelines for producing nondescript native goats. Furthermore, the findings emphasis how crucial it is to take stepwise regression models into account to precisely predict body weight in native goats of South Africa.

Acknowledgements

We extend our gratitude to Lepelle-Nkumpi Local Municipality farmers whose animals we were able to use for this research. Furthermore, we are grateful to Dr. S. Mogashoa, a former extension officer at Lepelle-Nkumbi Local Municipality, for helping us comprehend the traits of unique indigenous goats.

Conflict of interest

There are no conflicts of interest, according to the writers.

Declarations

We declare that this is our work.

Author details

Madumetja Cyril Mathapo*, Thobela Louis Tyasi, Kwena Mokoena,
Rankotsane Victoria Hlokoe and Madikadike Kagisho Molabe
Department of Agricultural Economics and Animal Production, University of Limpopo, Sovenga, Limpopo, South Africa

*Address all correspondence to: madumetja.mathapo@ul.ac.za

IntechOpen

References

[1] Sabbioni A, Beretti V, Superchi P, Ablondi M. Body weight estimation from body measures in Cornigliese sheep breed. Italian Journal of Animal Science. 2020;**19**(1):25-30

[2] Sandeep K, Dahiya SP, Malik ZS, Patil CS. Prediction of body weight from linear body measurements in sheep. Indian journal of. Animal Research. 2017;**52**(9):1263-1266

[3] Aytekin I, Eyduran E, Koksal K, Aksahan R, Keskin I. Prediction of fattening final live weight from somebody measurements and fattening period in young bulls of crossbred and exotic breeds using MARS data mining algorithm. Pakistan Journal of Zoology. 2018;**50**:189-195

[4] Rashijane LT, Mbazima VG, Tyasi TL. Prediction of body weight from linear body measurement traits of Boer goats raised at farm Tivolie, Limpopo Province, South Africa. American Journal of Animal and Veterinary Sciences. 2021;**16**:278-288

[5] Eyduran E, Zaborski D, Waheed A, Celik S, Karadas K, Grzesiak W. Comparison of the predictive capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the indigenous Beetal goat of Pakistan. Pakistan Journal of Zoology. 2017;**49**:257-265

[6] Sengul T, Celik S, Sengul O. Use of multivariate adaptive regression splines (MARS) for predicting parameters of breast meat in quails. Journal of Animal Plant Science. 2020; 27 86-793

[7] Yakubu A. Application of regression tree methodology in predicting the body weight of Uda sheep. Animal Science and Biotechnology. 2012;**33**(5):484-490

[8] Hlokoe VR, Mokoena K, Tyasi TL. Using multivariate adaptive regression splines and classification and regression tree data mining algorithms to predict body weight of Nguni cows. Journal of Applied Animal Research. 2022;**50**(1): 534-539

[9] Ali M, Eyduran E, Tariq MM, Tirink C, Abbas F, Bajwa MA, et al. Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai Sheep. Pakistan Journal of Zoology. 2015;**47**: 1579-1585

[10] Hasanah SH. Multivariate adaptive regression splines for modelling. The Student Status at Universitas Terbuka. 2021;**7**:51-58

[11] Khan MA, Tariq MM, Eyduran E, Tatliyer A, Rafeeq M, Abbas F, et al. Estimating body weight from several body measurements in Harnai sheep without multicollinearity problem. Journal of Animal and Plant Sciences. 2014;**24**:120-126

[12] Capricorn District Municipality (CDM). The agriculture statuses in the Capricorn District, Limpopo Province. Integral Development Plan. 2014/15

[13] Tyasi TL, Eyduran E, Celik S. Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa. Tropical Animal Health and Production. 2021;**53**(7):854-863

[14] Koc Y. Application of Regression Tree Method for Different Data from Animal Science. MSc Thesis. Turkey: Department of Animal Science, Agricultural Faculty, Igdir University; 2016. p. 58

[15] Olfaz M, Tırınk C, Önder H. Use of CART and CHAID Algorithms in Karayaka Sheep Breeding. Kafkas Univ Vet Fak Derg. 2019;**25**(1):105-110. DOI: 10.9775/kvfd.2018.20388

[16] Tirink C, Piwczynski D, Kolenda M, Onder H. Estimation of body weight based on biometric measurements by using random forest regression, support vector regression and CART algorithms. Animals. 2023;**13**(5):798

[17] Faraz A, Tirink C, Eyduran E, Waheed A, Tauqir NA, Nabeel MS, et al. Prediction of live body weight based on body measurements in thalli sheep under tropical conditions of Pakistan using CART and MARS. Tropical Animal Health and Production. 2021;**53**(301): 1-12. DOI: 10.1007/s11250-021-02748-6

[18] Berhe W. G: Relationship and prediction of bodyweight from morphometric traits in Maefur goat population in Tigray, Northern Ethiopia. Journal of Biomedical and Biostatistics. 2017;**8**(5):1-6. DOI: doi.org/10.4172/ 2155-6180.1000370

[19] Temoso O, Coleman M, Baker D, Morley P, Baleseng L, Makgekgenene A, et al. Using path analysis to predict bodyweight from body measurements of goats and sheep of communal rangelands in Botswana. South African Journal of Animal Science. 2017;**47**(6):854-863. DOI: doi.org/10.4314/sajas.v47i6.13

[20] Odadi WO. Using heart girth to estimate live weight of heifers (Bos indicus) in pastoral rangelands of northern Kenya. Livestock Research for Rural Development. 2018;**30**(1):1-9. Available from: http://www.lrrd.org/ lrrd30/1/wood30016.html

[21] Celik S, Yilmaz O. Comparison of different data mining algorithms for prediction of body weight from several morphological measurements in dogs. Journal of Animal and Plant Science. 2017;**27**(1):57-64

[22] Celik S. Comparing predictive performances of tree-based datamining algorithms and MARS algorithms in the prediction of live bodyweight from body traits in Pakistan goats. Pakistan Journal of Zoology. 2019;**51**(4):1447-1456. DOI: 10.17582/journal.pjz/ 2019.51.4.1447.1456

Chapter 4

Automated Data-Driven and Stochastic Imputation Method

Michal Koren and Or Peretz

Abstract

Machine learning algorithms may have difficulty processing datasets with missing values. Identifying and replacing missing values is necessary before modeling the prediction for missing data. However, studies have shown that uniformly compensating for missing values in a dataset is impossible, and no imputation technique fits all datasets. This study presents an Automated and data-driven Stochastic Imputer (ASI). The proposed ASI is based on automated distribution detection and estimation of the imputed value by stochastic sampling with controlled error probability. The significant advantage of this method is the use of a data-driven approximation ratio for the stochastic sampling, which bound the samples to be, at most, one standard deviation from the original distribution. The ASI performance was compared to traditional deterministic and stochastic imputation algorithms over seven datasets. The results showed that ASI succeeded in 61.5% of cases compared to other algorithms, and its performance can be improved by controlling the sampling error probability.

Keywords: imputation techniques, machine learning, multidimensional data, stochastic processes, artificial intelligence

1. Introduction

Currently, most Artificial Intelligence (AI) research and development is conducted in industry and academia, where vast quantities of information are generated daily. Furthermore, developing and deploying machine learning systems requires large amounts of data. Developing AI procedures is challenging since complete datasets (without missing values) are required. It is possible to remove missing values from a dataset or to impute them (e.g., replace them with comparative values), depending on the feature type [1]. There are established rules to decide which strategy to use for different types of missing values. Researchers have found that there is no one way to compensate for missing values in datasets. Also, specific datasets and types of missing data may respond well to specific strategies but not others [2, 3]. Data analysis can be affected by missing data by increasing or decreasing the value of specific categories when there is a large amount of missing data. In particular, missing data could affect machine learning (ML) algorithms and result in inaccurate and biased analysis.

Generally, numeric features are imputed using mean imputation (to avoid outliers and keep the data centralized), median, or mode. For categorical features, a "missing"

category is often added, or the most frequently occurring value assigned [4–6] is assigned. The use of ML algorithms and advanced statistical methods to complete missing values has been added to the conventional methods. KNN imputers [7–9] are commonly used to find neighbors and calculate values based on feature means. An additional imputation technique is a Multivariate Imputation by Chained Equations (MICE) algorithm [10]. MICE is a robust, informative method that imputes missing data in a dataset through an iterative series of predictive models [11, 12]. When examining more complex models, models with a higher learning capacity can complete missing values. Many deep learning architectures have been created to solve imputation challenges using the latent representation of the data in the hidden layers [13] and a combination of ML techniques [14], autoencoders [15], and multilayer perceptron [16]. In addition to the techniques discussed above, stochastic imputation techniques, such as regression models [17–19], add random errors to the target value and examine whether they correlate with the independent variable. Extrapolation and interpolation can be used to estimate unknown values by extending a known sequence of values or facts to unknown values [20, 21]. Conversely, non-deck techniques require substituting observations from "similar" units for each missing value [22, 23].

A new field of research is emerging in addition to machine learning techniques: automated machine learning (AutoML). ML problems can be solved more efficiently and effectively with AutoML [24, 25]. It requires search and optimization methods to find the best hyperparameters for a given problem [26]. It is challenging to impute the missing values in each feature, even with automatic methods [27]. This challenge may be due to high anomaly rates or data without patterns. As only some datasets have the same distribution and dependencies, human intervention is necessary to determine the appropriate value to impute.

The main challenge in the imputation process occurs among datasets with noisy distributions and anomalous values [28, 29]. When the distributions are inconsistent and contain many abnormal values, the imputation process becomes a challenge since all the statical measures are biased according to the behavior of the noises [30–32]. As part of stochastic processes and the use of randomization in algorithms, there is a need to allow a probability of failure. Since randomized decisions can change the entire process, analyzing the probability of the worst case of an algorithm is essential [33]. Moreover, it is useful to know the probability of whether the received answer is incorrect and handle it accordingly [34].

Three inequalities analyze the farthest value the random variable can be taken from its mean: (1) Markov inequality, which bounds the probability by the expected value of the variables [35]; (2) Chebyshev inequality, which bounds the probability by the variance [36, 37]; and (3) Chernoff inequality, which aims to achieve a binomial distribution (i.e., the sum of Bernoulli variables) that bounds the probability by exponential function [38, 39]. With the help of these inequalities, it is possible to develop random algorithms and control the result obtained by analyzing the correctness of the algorithm [40].

At times, complete independence can imply exponentially better bounds. Compared to Chebyshev, which only uses pairwise independence, Chernoff gives a tighter bound for deviation probability since it uses complete independence between the random variables. Although the Chernoff bound requires stronger assumptions, it is generally tighter than Markov and Chebyshev inequalities.

This study presents the Automated Stochastic Imputer (ASI), a new automated data-driven and stochastic method to impute numeric values in a dataset. It is based on the automated detection of distribution and estimation of the imputed value by

sampling with controlled error probability. The innovation of this method is the use of a data-driven approximation ratio based on the distribution measures and the determination of the number of samples required for an accurate estimation. Section 2 presents the method, its implementation, correctness, and computational complexity analysis. Section 3 describes the empirical study and a detailed scenario in which the results of the proposed method are compared to the existing imputation algorithm, with the results presented in Section 4. Last, Section 5 discusses the main conclusions and suggestions for future directions.

2. Automated Stochastic Imputer

This section presents and describes the ASI method. First, the definitions and distributions used in this study will be presented, followed by the method and description of its implementation. Last, the correctness of the method and the computational complexity will be detailed.

2.1 Definitions

The following are the definitions and methods employed in this study:

1. Chernoff bound [40] was used to bound the error probability. Let $X_1, \ldots, X_t$ be independent and identically distributed random variables with a range between zero to one, such that for all $1 \leq i \leq t$ it holds that $E[X_i] = \mu$. According to Chernoff bound, it holds that for any $0 < \epsilon < 1$, the probability of the sample's mean average (i.e., $\frac{1}{t}\sum_{i=1}^{t} X_i$) to be ϵ-far from the distribution mean is:

$$P\left[\left|\frac{1}{t}\sum_{i=1}^{t} X_i - \mu\right| \geq \epsilon\mu\right] \leq 2e^{-\frac{\mu t\epsilon^2}{3}} \tag{1}$$

2. Let $\mathbb{D}$ be a distribution, as listed in Appendix A. To determine whether each feature is indeed close enough to $\mathbb{D}$ distribution, a Kolmogorov-Smirnov test is performed [41]. The Kolmogorov-Smirnov test is a general nonparametric method. This test compares the empirical cumulative distribution functions of a sample with a postulated theoretical distribution. The Kolmogorov-Smirnov test is performed for each feature. In this case, the test interpretation is as follows:

H_o: The sample follows a $\mathbb{D}$ distribution.
H_1: The sample does not follow a $\mathbb{D}$ distribution.

2.2 Algorithm and implementation

Let $F = \{F_1, \ldots, F_m\}$ be a set of features in dataset D and let δ be the desired imputer failure probability, where $0 < \delta < 1$. First, the method is iterated over each feature $f \in F$ and normalizes the values into a range between zero to one by min-max normalization, i.e., each value of $v \in f$ is transformed to:

$$\frac{v - \min(f)}{\max(f) - \min(f)} \tag{2}$$

Next, the method estimates the distribution of f using the Kolmogorov-Smirnov test [41]. Let $\mathbb{D}(f)$ be the estimated distribution with a confidence level of α, and let f_μ,f_σ be the feature's expected value and standard deviation, respectively. The method defines q, the number of required samples necessary to estimate missing values, with a probability of $1 - \delta$, as:

$$q = \left\lceil -3\left(\frac{f_\mu}{f_\sigma}\right)\ln\left(\frac{\delta}{2}\right)\right\rceil \tag{3}$$

Let V_f be a set of missing values in f. The method iterates over each $u_i \in V_f$ and samples q independent and identically distributed values from $\mathbb{D}(f)$, denoted as $x_1, \ldots, x_q$. Last, the method imputes the missing values according to their average.

Automated Stochastic Imputer (D, α, δ)

a. $F \leftarrow \{F_1, \ldots, F_m\}$ set of features in D

b. For $f \in F$

- $\mathbb{D}(f) \leftarrow$ distribution of f with confidence level of α
- $q \leftarrow \left\lceil -3\left(\frac{f_\mu}{f_{\sigma^2}}\right)\ln\left(\frac{\delta}{2}\right)\right\rceil$
- $V_f \leftarrow$ set of missing values in f
- For $u_i \in V_f$:

  Sample $x_1, \ldots, x_q$ from $\mathbb{D}(f)$

  $u_i \leftarrow \frac{1}{q}\sum_{j=1}^{q} x_j$

c. Return D

2.3 Correctness

Let D be a dataset consisting of m features, denoted as $F = \{F_1, \ldots, F_m\}$. Initially, the method is iterated over the feature values and normalized by min-max normalization. For each value of $f \in F$, let $v \in f$ be transformed to:

$$\frac{v - \min(f)}{\max(f) - \min(f)} \tag{4}$$

Therefore, the dataset is normalized into a range between zero and one, preserving the original data distribution. For $v_a, v_b \in f$, such that $v_a \leq v_b$, it holds that:

$$v_a - \min(f) \leq v_b - \min(f) \tag{5}$$

$$\frac{v_a - \min(f)}{\max(f) - \min(f)} \leq \frac{v_b - \min(f)}{\max(f) - \min(f).} \tag{6}$$

Let $\mathbb{D}(f)$ be the detected distribution acquired by the Kolmogorov-Smirnov test with a confidence level of α, and let $x_1, \ldots, x_q \sim \mathbb{D}(f)$ be the independent and identical distributed random variables, such that $\forall 1 \leq i \leq q : E[x_i] = \mu_x$ and $V[x_i] = \sigma_x^2$. The average of all samples, denoted as $\overline{X}$, is defined as:

$$\overline{X} = \frac{1}{q}\sum_{i=1}^{q} x_i \tag{7}$$

Let μ, σ^2 be the expected value and the variance of $\overline{X}$, respectively. The method uses $\overline{X}$ as a single imputed value. By the linearity of expectation, it holds that the expected value of $\overline{X}$ is:

$$\mu = E\left[\overline{X}\right] = E\left[\frac{1}{q}\sum_{i=1}^{q} x_i\right] = \frac{1}{q}E\left[\sum_{i=1}^{q} x_i\right] = \frac{q \cdot E[x_i]}{q} = \mu_x \tag{8}$$

Given that all x_i are independently and identically distributed, it holds that:

$$\forall i \neq j : Var\left[x_i + x_j\right] = Var[x_i] + Var\left[x_j\right] \tag{9}$$

$$\sigma^2 = Var\left[\overline{X}\right] = Var\left[\frac{1}{q}\sum_{i=1}^{q} x_i\right] = \frac{1}{q^2}\sum_{i=1}^{q} Var[x_i] = \frac{q \cdot Var[x_i]}{q^2} = \frac{\sigma_x^2}{q} \tag{10}$$

The probability error of each sample to be ϵ-far from the expected value of the feature can be bound by Chernoff inequality:

$$P\left[\left|\frac{1}{q}\sum_{i=1}^{t} X_i - \mu\right| \geq \epsilon\mu\right] \leq 2e^{-\frac{\mu q \epsilon^2}{3}} \tag{11}$$

Let $\epsilon = \frac{\sigma_x}{\mu_x}$ be the approximation factor. By bounding the failure probability with δ, a lower bound will be received for the number of samples required:

$$P\left[\left|\frac{1}{q}\sum_{i=1}^{t} X_i - \mu\right| \geq \left(\frac{\sigma_x}{\mu_x}\right)\mu_x\right] \leq 2e^{-\frac{q \cdot \mu_x \left(\frac{\sigma_x}{\mu_x}\right)^2}{3}} < \delta \tag{12}$$

Thus, the probability of $\overline{X}$ to be ϵ-far from the expected value is equal to its probability to be the distance of, at most, one standard deviation:

$$P\left[\left|\frac{1}{q}\sum_{i=1}^{t} X_i - \mu\right| \geq \left(\frac{\sigma_x}{\mu_x}\right)\mu_x\right] = P\left[\left|\frac{1}{q}\sum_{i=1}^{t} X_i - \mu\right| \geq \sigma_x\right] \leq 2e^{-\frac{q \cdot \mu_x \left(\frac{\sigma_x}{\mu_x}\right)^2}{3}} < \delta \tag{13}$$

Applying algebraic simplification:

$$q > -\frac{3\mu_x \ln\left(\frac{\delta}{2}\right)}{\sigma_x^2} = -\frac{3\mu_x}{\sigma_x^2} \ln\left(\frac{\delta}{2}\right) = -3\left(\frac{\mu_x}{\sigma_x^2}\right) \ln\left(\frac{\delta}{2}\right) \tag{14}$$

Since all the dataset distributions are normalized into the range [0,1], its expected value and standard deviation are non-negative values. Given that $0 < \delta < 1$, the number of samples (i.e., q) is a positive number:

$$0 < \delta < 1 \Rightarrow \ln\left(\frac{\delta}{2}\right) < 0 \tag{15}$$

$$q > -3\left(\frac{\mu_x}{\sigma_x^2}\right) \ln\left(\frac{\delta}{2}\right) > 0 \tag{16}$$

Therefore, it can be concluded that the total number of samples required to estimate an imputed value depends on the ratio between the expectation and the distribution variance.

Last, for a more accurate analysis of the imputed values, the following is the proof that all values will be typically distributed around the feature mean: Let $V_f = (m_1, \ldots, m_M)$ be a set of M missing values in feature f. For all $1 \le i \le q$, it holds that $x_i \sim \mathbb{D}(f)$, and by the imputation technique:

$$\forall 1 \le j \le M : m_j = \frac{1}{q} \sum_{i=1}^{q} x_i \tag{17}$$

As long as $M \leftarrow \infty$, by the central limit theorem, the imputed values will be normally distributed:

$$\overline{M} = \frac{1}{M} \sum_{j=1}^{M} m_i \sim N\left(\mu_x, \frac{\sigma_x^2}{M}\right) \tag{18}$$

2.4 Computational complexity

Let m be the number of features in a dataset with n records. For each feature with at least one missing value, the method iterates over the missing values. Therefore, the maximum number of such features is m. The upper bound of missing values in each feature is n. Let q be the number of samples required to estimate the missing value of each iteration. Thus, sampling and averaging their results require $2q$ operations, and the total complexity time is:

$$O(m \cdot n \cdot (2q)) \approx O(mnq) \tag{19}$$

In a standard dataset, there are fewer features and records, i.e., $m \le n$, and the complexity running time can be bound by:

$$O(mnq) \approx O(n^2 q) \tag{20}$$

3. Empirical study

3.1 Data sources

Seven datasets were compared to examine the ASI method:

1. *Fetal health* [42]—a medical dataset that aims to prevent child and maternal mortality. The dataset consists of 2126 observations over 21 features and a target variable with three values: normal, suspect, or pathological.

2. *Students' academic success* [43]—this dataset contains information about students with different undergraduate degrees from higher education institutions. It includes information on 4424 students' enrollment and academic performance over 36 features. The target variable has three options: graduate, dropout, or enrolled.

3. *Heart failure* [44]—a dataset with a total of 299 patients who experienced heart failure. It contains 12 clinical features and a Boolean target variable representing whether the patient had heart failure.

4. *Diabetes* [45]—a dataset of 768 diabetic and non-diabetic women. It consists of six medical features, two demographic variables and a Boolean target variable.

5. *Haberman's survival* [46]—a dataset collected from 1958 to 1970 that includes details on 306 survivors after breast cancer surgery. The dataset and study were conducted at the University of Chicago's Billings Hospital.

6. *Breast cancer* [47]—a dataset with a total of 30 features extracted from digitized images of diagnosed breast cancer of 568 patients. The target variable indicates whether the "mass" diagnosis was benign or malignant.

7. *Bank* [48]—over 45,210 observations of direct marketing campaigns by Portuguese banks are included in the dataset, including seven numerical (continuous) and nine categorical features. The Boolean target variable indicates whether the client subscribed to a term deposit or not.

3.2 Experiment procedure

For each dataset described in Section 3.1, the following parameters were defined and used:

1. At the beginning of each experiment, 25% of the data were randomly chosen, removed, and stored as a version of the original dataset to compare the imputation techniques.

2. For the distribution estimation using the Kolmogorov-Smirnov test, a total of 100 samples with a significance level of $\alpha = 0.05$ were used.

3. The results of the proposed ASI method were compared to the existing algorithms implemented in scikit-learn [49] as follows:

    i. *KNN imputer*—imputation for completing missing values using nearest neighbors' algorithm.

    ii. *Iterative imputer*—multivariate imputation by chained equations (MICE) [10]. As this imputer has stochastic phases, a random state was set equal to zero in all runs presented in this study.

4. For each comparison, the values $\delta = 0.1, 0.05, 0.01$ were examined for the probability of failure in estimation.

3.3 Use case: the heart failure dataset

To simplify the demonstration, the heart failure dataset, consisting of 12 features over 299 observations, was chosen. The feature names were denoted as F_i for all $1 \leq i \leq 12$. A broader comparison, including higher-dimensional datasets, can be found in Section 4. For the scenario demonstration, 56 arbitrary values were randomly removed (≈25%) and selected from 5 arbitrary features. Their value was then stored to compare to the proposed method. **Table 1** presents the initial number of missing values for each feature and the automated distribution detected by the Kolmogorov-Smirnov test.

Once the distributions were detected, the method calculated the number of samples required for each feature (i.e., q) by the ratio between its expected value and variance. **Table 2** compares the values of q and the performance of the ASI method, the KNN imputer, and the MICE imputer. Since the proposed method input an upper bound on the probability of failure (i.e., δ), between $\delta = 0.1, 0.05, 0.01$ was compared. As the probability of error was smaller, the number of samples required to estimate the missing value increased, and the performance of the ASI method improved. For example, feature 5 for probability 0.1 yielded 40% success; 50% success was obtained for probability 0.05, and when the probability of error was bounded by 1% (i.e., 0.01), it increased to 60% success compared to the other algorithms.

For each compared delta (i.e., the probability of wrong estimation), the percentage of total successes of each imputation algorithm was calculated. The results are presented in **Figure 1**. The smaller the probability of error, the more samples from the distribution were required to estimate the imputed value. Accordingly, the performance of the ASI algorithm increased. For a 10% chance of wrong estimation, the proposed method succeeded in 54% of the cases, compared to 27% of MICE and 19%

Feature	Missing values	Expectation	Standard deviation	Distribution
F1	15	0.438	0.497	Normal
F4	10	38.059	11.725	Exponential
F5	10	0.416	0.494	Exponential
F6	11	1.372	0.995	Gamma
F8	13	129.272	78.065	Beta

Table 1.
Number of missing values in each feature and its detected distribution.

Feature	q	ASI	KNN	MICE
$\delta = 0.1$				
F1	8	11 (73%)	1 (7%)	3 (20%)
F4	8	5 (50%)	2 (20%)	3 (30%)
F5	30	4 (40%)	4 (40%)	2 (20%)
F6	13	7 (64%)	1 (9%)	3 (27%)
F8	15	5 (38%)	3 (24%)	5 (38%)
$\delta = 0.05$				
F1	10	12 (80%)	1 (7%)	2 (13%)
F4	10	6 (60%)	2 (20%)	2 (20%)
F5	36	5 (50%)	3 (30%)	2 (20%)
F6	16	7 (64%)	2 (18%)	2 (18%)
F8	19	6 (46%)	3 (23%)	4 (31%)
$\delta = 0.01$				
F1	15	12 (80%)	1 (7%)	2 (13%)
F4	14	7 (70%)	1 (10%)	2 (20%)
F5	52	6 (60%)	2 (20%)	2 (20%)
F6	22	7 (64%)	2 (18%)	2 (18%)
F8	27	10 (77%)	1 (8%)	2 (15%)

Notes. *The data is represented as A(B), where A is the count of successes, and B is their percentages. q is the number of samples required for ASI.*

Table 2.
A comparison between the performance of the three imputation algorithms.

Figure 1.
Performance of three imputation algorithms compared by failure probability.

of KNN. When the chance of error was only 1%, the proposed method succeeded in 71% of cases, representing a significant increase compared to KNN's 12%.

4. Results

To illustrate the ASI method, seven datasets (for details, see Section 3.1) were compared with the parameters defined in Section 3.2. For each dataset, the four algorithms were evaluated, and their performances were compared. First, the

Dataset	Missing values	ASI	KNN	MICE
Fetal health	425	210 (49%)	98 (23%)	117 (28%)
Students	94	49 (52%)	27 (29%)	18 (19%)
Diabetes	152	84 (55%)	42 (28%)	26 (17%)
Heart failure	59	42 (71%)	7 (12%)	10 (17%)
Haberman survival	26	20 (77%)	1 (4%)	5 (19%)
Cancer	113	78 (69%)	16 (14%)	19 (17%)
Bank	13,276	7687 (58%)	2276 (17%)	3313 (25%)

Notes. *The data is represented as A(B), where A is the count of successes, and B is their percentages. The comparison used* $\delta = 0.05$ *for the ASI.*

Table 3.
Performance of the three imputation algorithms compared by dataset.

comparison results and analysis will be described. Then, the sensitivity analysis of the proposed method will be presented. **Table 3** presents the total missing value for each dataset and the success percentages for each imputer.

The ASI method achieved the highest success rates for all seven tested datasets compared to the KNN and MICE imputers. An extreme case was demonstrated in the Haberman survival dataset, for which a 77% success rate was recorded for the proposed method (i.e., 20 correct answers out of 26). Upon further analysis, this dataset consisted of discrete value features. Therefore, the suitability of the proposed method for discrete versus continuous values should be further examined in future studies.

To examine the error distributions, the errors obtained for each algorithm were calculated by subtracting the original value from the imputed value. These errors were then averaged for each dataset. Since this comparison was between the average values (and not distributions), polynomial interpolation was performed [50]. The results are shown in **Figure 2**.

Two conclusions can be drawn from **Figure 2**. First, there is a high level of agreement between the three algorithms; that is, the ASI algorithm behaved similarly (distributive)

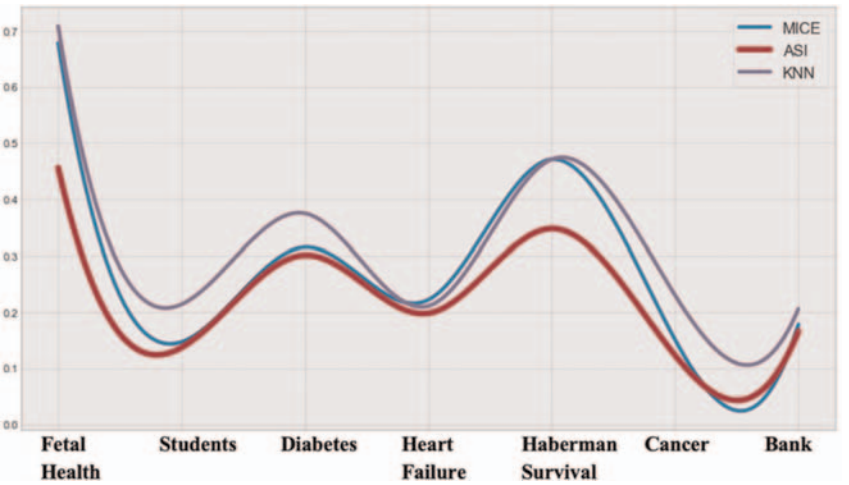

Figure 2.
Polynomial interpolation of success rates of the three compared algorithms.

Dataset	δ								
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Fetal health	35.4	34.2	32.8	32.5	33.7	32.1	33.2	31.6	28.3
Students	40.2	37.1	40.1	39.1	37.2	42.3	37.1	40.2	39.4
Diabetes	38.2	36.9	35.9	35.3	33.6	34.4	38.2	39.6	30.3
Heart failure	43.1	42.7	46.5	41.3	42.1	39.6	39.1	39.4	39.9
Haberman survival	70.1	85.2	77.3	78.4	75.2	75.1	71.6	69.6	63.9
Cancer	74.3	75.2	73.8	73.4	74.1	71.6	70.4	70.9	67.2
Bank	49.8	47.4	46.1	44.6	42.6	41.6	40.3	40.8	40.2

Table 4.
Performance of the ASI algorithm, compared by probability of failure.

to the KNN and MICE algorithms. Second, the ASI algorithm presented a low percentage of all the errors examined compared to the others. However, the ASI algorithm was probabilistic and dependent on input parameters that controlled the probability of failure. Thus, further analysis and comparison of different probabilities is required.

4.1 Sensitivity analysis

This section compares the performance of the ASI algorithm regarding different values of failure probabilities. The proposed algorithm was evaluated using each dataset with probabilities ranging from 0.1 to 0.9 with jumps of 0.1. Then, the percentage of successes the algorithm made was calculated and compared to the actual value. The results are shown in **Table 4**. Since the goal was to examine the algorithm's behavior and compare the success percentages, polynomial interpolation was performed [50], as presented in **Figure 3**.

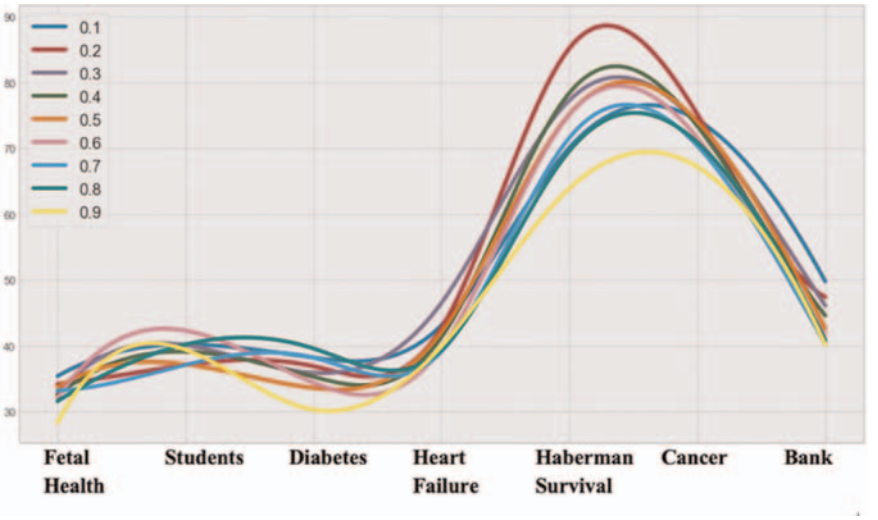

Figure 3.
A comparison between success rates of ASI over different probabilities.

In the sensitivity analysis of the ASI algorithm, it can be seen that each dataset behaved similarly for different probabilities. This conclusion indicates the consistency of the proposed method. It can also be concluded that noisier results were obtained for extreme probabilities (low and high) due to more substantial constraints on the method's inputs (e.g., probabilities 0.9 and 0.2). One non-intuitive result was a probability of failure of 0.1, whose values appeared to be "outliers" from the rule. This result might be due to the randomness of the algorithm since, for each probability, the sampled data changed according to the method. However, this specific result was unusual because the other probabilities agreed on the general behavior of the proposed algorithm.

5. Discussion and conclusions

This study proposed a new automated data-driven and stochastic imputation method, ASI, to complete missing values in a dataset. The ASI is based on automated distribution detection and estimation of the imputed value by sampling with controlled error probability. This study's innovation was the use of a data-driven approximation ratio based on the distribution measures and determination of the number of samples required for an accurate estimation. Thus, the imputer successfully bound the distance between the imputed value and the original expectation by, at most, one standard deviation. The following are the main conclusions:

1. The ASI method succeeded in imputing the missing values in 61.5% of the cases, compared to the deterministic KNN and stochastic MICE algorithms. According to the analysis, there was a slightly difference in the performance of the proposed algorithm on features with discrete and continuous values.

2. The number of samples required to estimate missing values increased as the error probability decreased. Thus, more samples yielded better estimation performance, although it increased the runtime complexity. For example, in the heart failure dataset (Section 3.3), the first feature required eight samples to yield a success rate of 73%, while the use of 15 samples improved the success rate to 80%.

3. A sensitivity analysis of the ASI algorithm on different probabilities found consistency in its performance on seven datasets. It can be concluded that the correctness of the proposed method is tight and presents an accurate imputation of missing values in a given distribution.

Results are known to be affected by the data quality, especially when considering data imputation and stochastic models. Failures may result from a lack of data or incorrect adjustment of the parameters, as mentioned in this method. Future studies should address two main issues presented in this study. First, the performance of ASI over continuous and discrete random variables should be explored, which may be done by examining different values for the parameters or an extension of the distribution exploration. Second, this study assumed a confidence level of $\alpha = 0.05$ for distribution testing using the Kolmogorov-Smirnov test. Different significance levels may yield different results of distributions and, thus, the ASI algorithm's results.

Conflict of interest

The authors declare no conflict of interest.

Appendix

See **Table A1**.

Type	Distribution	Parameters	Expectation	Variance
Discrete	Uniform	$U(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
	Binomial	$Bin(n,p)$	np	$np(1-p)$
	Geometric	$G(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
	Hypergeometric	$HG(N,D,n)$	$\frac{nD}{N}$	$\frac{nD}{N}\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right)$
	Poisson	$Pois(\lambda)$	λ	λ
	Negative Binomial	$NB(n,p)$	$\frac{n(1-p)}{p}$	$\frac{n(1-p)}{p^2}$
Continuous	Uniform	$U_c(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	Exponential	$\exp(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
	Normal	$N(\mu,\sigma^2)$	μ	σ^2
	Gamma	$Gamma(n,\lambda)$	$\frac{n}{\lambda}$	$\frac{1}{\lambda^2}$
	Beta	$Beta(a,b)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$
	Chi2	$\chi^2(k)$	k	$2k$
	F	$F(m,n)$	$\frac{n}{n-2}$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$
	T	$T(m)$	0	$\frac{m}{m-2}$

Table A1.
List of distributions to determine the feature.

Author details

Michal Koren* and Or Peretz
School of Industrial Engineering and Management, Shenkar—Engineering, Design, Art, Ramat-Gan, Israel

*Address all correspondence to: michal.koren@shenkar.ac.il

References

[1] Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2010;**59**(10):1087-1091. DOI: 10.1016/j.jclinepi.2006.01.014

[2] Newman DA. Missing data: Five practical guidelines. Organizational Research Methods. 2014;**17**(4): 372-411. DOI: 10.1177/1094428114548590

[3] Salgado CM, Azevedo C, Proença H, Vieira SM. Missing data. In: Secondary Analysis of Electronic Health Records. MIT Critical Data. Cham: Springer; 2016. pp. 143-162

[4] Akande O, Li F, Reiter J. An empirical comparison of multiple imputation methods for categorical data. The American Statistician. 2017;**71**(2): 162-170. DOI: 10.1080/00031305.2016.1277158

[5] Finch WH. Imputation methods for missing categorical questionnaire data: A comparison of approaches. Journal of Data Science. 2010;**8**(3):361-378

[6] Schuckers M, Lopez M, Macdonald B. Estimation of player aging curves using regression and imputation. Annals of Operations Research. 2023;**325**:681-699. DOI: 10.1007/s10479-022-05127-y

[7] Koren M, Koren O, Peretz O. Weighted distance classification method based on data intelligence. Expert Systems. 2023;**41**(2):e13486. DOI: 10.1111/exsy.13486

[8] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;**17**(6): 520-525. DOI: 10.1093/bioinformatics/17.6.520

[9] Zhang S. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software. 2012;**85**(11): 2541-2552. DOI: 10.1016/j.jss.2012.05.073

[10] van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;**45**:1-67. DOI: 10.18637/jss.v045.i03

[11] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;**20**(1): 40-49. DOI: 10.1002/mpr.329

[12] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011; **30**(4):377-399. DOI: 10.1002/sim.4067

[13] Biessmann F, Rukat T, Schmidt P, Naidu P, Schelter S, Taptunov A, et al. DataWig: Missing value imputation for tables. Journal of Machine Learning Research. 2019;**20**(175):1-6

[14] Phung S, Kumar A, Kim J. A deep learning technique for imputing missing healthcare data. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 23-27 July 2019; Berlin, Germany. IEEE; 2019. pp. 6513-6516. DOI: 10.1109/EMBC.2019.8856760

[15] Duan Y, Lv Y, Liu YL, Wang FY. An efficient realization of deep learning for traffic data imputation. Transportation Research Part C: Emerging Technologies. 2016;**72**:168-181. DOI: 10.1016/j.trc.2016.09.015

[16] Lin WC, Tsai CF, Zhong JR. Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems. 2022;**239**:Article 108079. DOI: 10.1016/j.knosys.2021.108079

[17] Gold MS, Bentler PM. Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. Structural Equation Modeling. 2000;**7**(3):319-355. DOI: 10.1207/S15328007SEM0703_1

[18] Juan AA, Keenan P, Martí R, McGarraghy S, Panadero J, Carroll P, et al. A review of the role of heuristics in stochastic optimisation: From metaheuristics to learnheuristics. Annals of Operations Research. 2023;**320**(2): 831-861. DOI: 10.1007/s10479-021-04142-9

[19] Shehadeh KS, Padman R. Stochastic optimization approaches for elective surgery scheduling with downstream capacity constraints: Models, challenges, and opportunities. Computers & Operations Research. 2022;**137**:105523. DOI: 10.1016/j.cor.2021.105523

[20] Raja K, Arasu GT, Nair CS. Imputation framework for missing values. International Journal of Computer Trends and Technology. 2012; **3**(2):215-219

[21] Soeffker N, Ulmer MW, Mattfeld DC. Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review. European Journal of Operational Research. 2022;**298**(3): 801-820. DOI: 10.1016/j.ejor.2021.07.014

[22] Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. International Statistical Review. 2010;**78**(1):40-64. DOI: 10.1111/j.1751-5823.2010.00103.x

[23] Kim JK, Fuller W. Fractional hot deck imputation. Biometrika. 2004; **91**(3):559-578. DOI: 10.1093/biomet/91.3.559

[24] Wu Y, Xi X, He J. AFGSL: Automatic feature generation based on graph structure learning. Knowledge-Based Systems. 2022;**238**:Article 107835. DOI: 10.1016/j.knosys.2021.107835

[25] Yao Q, Wang M, Chen Y, Dai W, Li Y-F, Wei-Wei T, et al. Taking human out of learning applications: A survey on automated machine learning. 2018; arXiv:1810.13306. DOI: 10.48550/arXiv.1810.13306

[26] He X, Zhao K, Chu X. AutoML: A survey of the state-of-the-art. Knowledge-Based Systems. 2021;**212**: Article 106622. DOI: 10.1016/j.knosys.2020.106622

[27] Krishnan S, Franklin MJ, Goldberg K, Wu E. Boostclean: Automated error detection and repair for machine learning. 2017;arXiv: 1711.01299. DOI: 10.48550/arXiv.1711.01299

[28] Kenward MG, Carpenter J. Multiple imputation: Current perspectives. Statistical Methods in Medical Research. 2007;**16**(3):199-218. DOI: 10.1177/0962280206075304

[29] Schafer JL. Multiple imputation: A primer. Statistical Methods in Medical Research. 1999;**8**(1):3-15. DOI: 10.1177/096228029900800102

[30] Carpenter JR, Bartlett JW, Morris TP, Wood AM, Quartagno M, Kenward MG. Multiple Imputation and its Application. 2nd ed. Hoboken: John Wiley & Sons; 2023. 444 p

[31] Koren O, Koren M, Peretz O. A procedure for anomaly detection and

analysis. Engineering Applications of Artificial Intelligence. 2023;**117**:105503. DOI: 10.1016/j.engappai.2022.105503

[32] Ozkan H, Pelvan OS, Kozat SS. Data imputation through the identification of local anomalies. IEEE Transactions on Neural Networks and Learning Systems. 2015;**26**(10):2381-2395. DOI: 10.1109/TNNLS.2014.2382606

[33] Motwani R, Raghavan P. Randomized algorithms. ACM Computing Surveys. 1996;**28**(1):33-37

[34] Karp RM. An introduction to randomized algorithms. Discrete Applied Mathematics. 1991;**34**(1–3): 165-201. DOI: 10.1016/0166-218X(91)90086-C

[35] Cohen JE. Markov's inequality and Chebyshev's inequality for tail probabilities: A sharper image. The American Statistician. 2015;**69**(1):5-7. DOI: 10.1080/00031305.2014.975842

[36] Navarro J. A very simple proof of the multivariate Chebyshev's inequality. Communications in Statistics - Theory and Methods. 2016;**45**(12): 3458-3463. DOI: 10.1080/03610926.2013.873135

[37] Ogasawara H. The multivariate Markov and multiple Chebyshev inequalities. Communications in Statistics - Theory and Methods. 2020; **49**(2):441-453. DOI: 10.1080/03610926.2018.1543772

[38] Klaassen CA. On an inequality of Chernoff. Annals of Probability. 1985; **13**(3):966-974

[39] Rao BP, Sreehari M. Chernoff-type inequality and variance bounds. Journal of Statistical Planning and Inference. 199? 44 ?):325-335. DOI: 10.1016/S03?8-?758(97)00031-1

[40] Hwang CR, Sheu SJ. A generalization of Chernoff inequality via stochastic analysis. Probability Theory and Related Fields. 1987;**75**(1):149-157. DOI: 10.1007/BF00320088

[41] Massey FJ Jr. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association. 1951;**46**(253):68-78. DOI: 10.1080/01621459.1951.10500769

[42] Dua D, Graff C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science; 2019

[43] Realinho V, Martins MV, Machado J, Baptista LMT. Predict students' dropout and academic success data set. UCI Machine Learning Repository. 2021. DOI: 10.24432/C5MC89

[44] Chicco D, Giuseppe J. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making. 2020;**20**(1):1-16. DOI: 10.1186/s12911-020-1023-5. [Article ID: 16]

[45] Kahn M. Diabetes data set. UCI Machine Learning Repository. 1994. DOI: 10.24432/C5T59G. Available from: https://archive.ics.uci.edu/ml/datasets.php

[46] Haberman S. Haberman's survival data set. UCI Machine Learning Repository. 1999. DOI: 10.24432/C5XK51. Available from: https://archive.ics.uci.edu/ml/datasets.php

[47] Wolberg WH, Street WN, Mangasarian OL. Breast cancer Wisconsin (diagnostic). UCI Machine Learning Repository. 1995. DOI: 10.24432/C5DW2B. Available

from: https://archive.ics.uci.edu/ml/datasets.php

[48] Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 2014;**62**:22-31. DOI: 10.1016/j.dss.2014.03.001

[49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. Journal of Machine Learning Research. 2011;**12**:2825-2830

[50] Nakamura S. Numerical Analysis and Graphic Visualization with MATLAB. New York: Prentice-Hall, Inc.; 1995

Chapter 5

Exploring Feature Partitioning Methods for Data Mining Applications

Aditya Kumar and Jainath Yadav

Abstract

Feature partitioning is a fundamental concept in machine learning and data mining, offering a crucial framework for data representation, classification, and predictive modeling. This chapter delves into the multifaceted domain of feature partitioning, exploring the methodologies, techniques, and applications that drive this field. Feature partitioning methods range from random-based approaches to pattern-based, clustering-based, performance-based, and optimization-based techniques. The chapter provides a comprehensive overview of these methods, discussing their strengths, weaknesses, and suitability for various tasks. Furthermore, it analyzes the comparative performance of these methods, emphasizing their impact on classification accuracy. In addition to this evaluation, the chapter highlights the associated issues, challenges, and opportunities in the domain of multiview ensemble learning, offering a broader perspective on its future development. As a versatile concept with applications in diverse fields, feature partitioning has a crucial function in enhancing the quality and interpretability of machine learning models. This chapter serves as an excellent source for scholars, practitioners, and students seeking a deeper understanding of feature partitioning and its significance in modern data mining applications.

Keywords: multiview ensemble learning, feature set partitioning, views construction, ensemble learning, classification

1. Introduction

In today's data-driven world, the explosive growth of data has reshaped the landscape of machine learning and data analysis. Extracting meaningful insights from high-dimensional datasets is often a challenging endeavor, and traditional approaches have encountered limitations in managing the deluge of information. To address these challenges, multiview learning has emerged as a powerful paradigm, offering promising solutions to enhance the accuracy and robustness of predictive models. At the heart of multiview learning lies the fundamental concept of "Feature Partitioning" [1, 2].

1.1 The role of feature partitioning

Feature partitioning is a technique designed to address the curse of dimensionality and enhance the performance of machine learning models on high-dimensional data. It is a process of breaking down the feature space into smaller, more manageable subsets or "views" [3, 4]. Each view represents a distinct and coherent set of features that captures specific aspects of the data. By partitioning the feature space into multiple views, feature partitioning aims to:

- Reduce the dimensionality of the data, making it more manageable for machine learning algorithms.
- Uncover unique patterns and relationships within the data by analyzing each view individually.
- Enable the development of specialized models for each view, resulting in a comprehensive ensemble approach.
- Mitigate issues related to data sparsity and overfitting, enhancing generalization and accuracy of machine learning models.

Feature partitioning is not limited to any specific domain; it is a versatile strategy applicable across various fields. In genomics, it can aid in identifying relevant genetic markers for disease diagnosis. In finance, it can enhance the prediction of market trends. In natural language processing, it can assist in language understanding and translation. Its potential applications are as diverse as the datasets themselves.

1.2 The significance of feature partitioning

Feature partitioning is a pivotal technique in multiview ensemble learning (MEL), wherein high-dimensional datasets are divided into subsets of features or "views" [5, 6] as shown in **Figure 1**. Each view provides a distinct perspective on the underlying data, allowing for the capture of unique patterns and relationships within the information. By partitioning the feature space, feature partitioning methods aim to strike a balance between informativeness and redundancy, offering an effective means to mitigate the curse of dimensionality. This process facilitates the creation of diverse and complementary feature subsets that can be leveraged to build more accurate predictive models [3, 7, 8]. The significance of feature partitioning transcends numerous domains, from healthcare and finance to natural language processing and image recognition. In healthcare, for instance, it can help identify relevant biomarkers for disease diagnosis or treatment. In finance, it can improve the prediction of market trends and investment decisions. The applications of feature partitioning are as diverse as the datasets themselves, and its potential to boost classification accuracy and interpretability is paramount.

1.3 The objectives of this chapter

This book chapter is dedicated to exploring the realm of feature partitioning in multiview learning. It aims to provide an in-depth understanding of feature

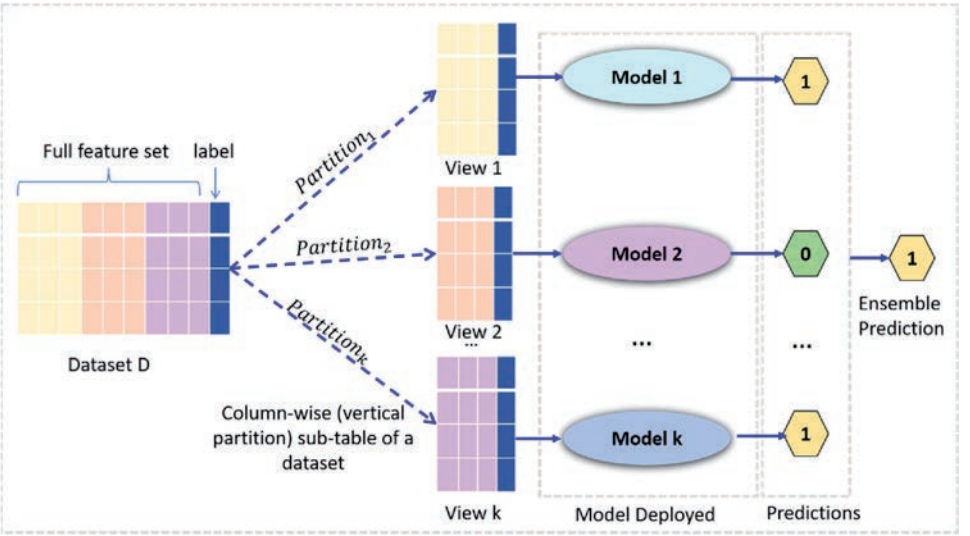

Figure 1.
Feature set partitioning-based multiview ensemble learning.

partitioning methods, their applications, and the comparative study of their performance. The objectives of this chapter include:

- An introduction to various feature partitioning methods, categorizing them into distinct groups.
- A comparative study that delves into the robustness and effectiveness of feature partitioning methods.
- An examination of common challenges and issues associated with feature partitioning in multiview ensemble learning.
- A look into future perspectives and potential advancements in the field of feature partitioning, including the incorporation of deep learning methods, addressing noisy and imbalanced data, and exploring applications in diverse domains.
- A comprehensive summary of findings and insights, underscoring the significance of feature partitioning in modern machine learning.

By achieving these objectives, this chapter aspires to provide readers with a solid foundation for understanding the principles, challenges, and potential applications of feature partitioning. It is our hope that this exploration will serve as a valuable resource for researchers, practitioners, and enthusiasts in the field of multiview learning and machine learning at large.

1.4 Structure of the chapter

The subsequent sections of this chapter will delve into feature partitioning methods, offer a comparative study, examine common challenges, and explore future prospects in more detail, providing a comprehensive view of this dynamic field.

2. Feature partitioning methods

In this section, we will delve into various feature partitioning methods, each offering distinct principles and applications. These methods play a crucial role in breaking down high-dimensional feature spaces into more manageable views, allowing for enhanced data analysis and improved machine learning model performance. We categorize these methods into six categories: random-based, pattern-based, clustering-based, performance-based, optimization-based, and ensemble-based (as shown in **Figure 2**). Let us explore each category in detail.

2.1 Random-based methods

Principles: Random-based feature partitioning methods are straightforward and operate without any specific patterns or guidelines. They allocate features to different views randomly to create diversity among the views. The primary goal is to reduce the correlation between views. Some random-based feature partitioning methods are given below:

- *Random Split* [9]*:* This method randomly divides the feature set into multiple views. The absence of any structured criteria for partitioning aims to provide each view with a different subset of features. It is commonly used as a baseline approach for comparative studies.

- *Attribute Bagging* [10]*:* Attribute Bagging introduces an element of optimization into random partitioning. It starts by conducting random feature selection experiments multiple times to identify the best subset size of attributes that leads to improved classification accuracy. Once the optimal subset size is determined, features are randomly partitioned into views based on the selected size.

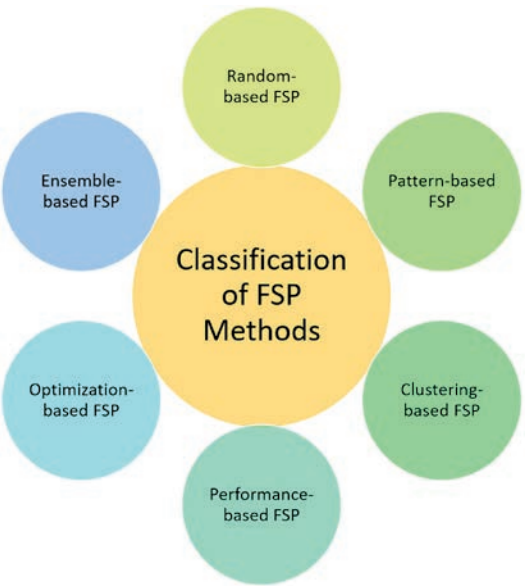

Figure 2.
Feature partitioning methods.

2.2 Pattern-based methods

Principles: Pattern-based feature partitioning methods allocate features to views based on specific patterns, themes, or sequences. These patterns can be predefined or extracted from the dataset itself, providing a more structured approach to partitioning. The aim is to create views with meaningful and distinct feature subsets. Some pattern-based feature partitioning methods are given below:

- *Bell Triangle-Based FSP* [11]*:* This method leverages Bell Triangles, a mathematical concept, to guide the partitioning of features. The structure of the Bell Triangle influences how features are divided into views.

- *Frequent Itemsets-Based FSP* [12]*:* Frequent Itemset mining techniques are used to discover patterns within the feature set. These patterns then dictate how the features are partitioned into different views.

- *Theme-Based FSP* [13]*:* In this approach, features are grouped into different views based on predefined themes or categories. The partitioning is driven by the theme assigned to each feature.

- *Ferrers Diagram-Based FSP* [11]*:* Ferrers diagrams, a mathematical concept, are employed to create partitions based on the structure of these diagrams.

- *Supervised FSP* [14]*:* Supervised FSP methods incorporate zig-zag pattern to guide the allocation of features to different views. By leveraging the information from the feature subset, it aims to create views that are tailored to specific classification tasks.

2.3 Clustering-based methods

Principles: Clustering-based feature partitioning methods focus on grouping features based on their inherent similarities or differences. These groups of attributes can be either homogeneous or heterogeneous, and the primary goal is to form views with coherent sets of features. Some clustering-based feature partitioning methods are given below:

- *Graph Coloring-Based FSP Method* [15, 16]*:* In this method, features are assigned to different clusters using graph coloring techniques. Each cluster represents a view, and the partitioning aims to minimize the similarity among features within the same cluster (intra-cluster similarity) while maximizing the dissimilarity among different clusters (inter-cluster dissimilarity).

- *Attribute Clustering* [17]*:* Attribute Clustering employs the k-means clustering technique to group features into clusters. Each cluster serves as a distinct view. The goal is to create views with features that are closely related to each other.

- *Collaboration Graph-Based FSP* [18]*:* Collaboration Graph-Based methods rely on collaborative relationships among features. Features that have collaborative connections are grouped in the same cluster, forming coherent views.

- *Minimum Spanning Tree-Based Feature Grouping* [19]*:* This method uses minimum spanning tree algorithms to group features into clusters. The structure of the spanning tree influences the partitioning of features into different views.

2.4 Performance-based methods

Principles: Performance-based feature partitioning methods aim to improve ensemble classification accuracy by conducting experiments multiple times. The iterative approach is used to refine the allocation of features to different views. Some performance-based feature partitioning methods are given below:

- *Optimal Feature Set Partitioning (OFSP) Technique* [3]*:* This technique repeatedly conducts experiments to identify the optimal feature partitioning that leads to enhanced ensemble classification accuracy. It fine-tunes the partitioning based on the results of these experiments.

- *Rough Set-Based OFSP* [20]*:* Leveraging rough set theory, this method optimizes feature partitioning by considering the rough boundaries of decision regions. The objective is to achieve high accuracy in ensemble classification.

2.5 Optimization-based methods

Principles: Optimization-based feature partitioning methods focus on optimizing evaluation criteria for the machine learning framework. These methods often employ metaheuristic optimization approaches to find the optimal feature partitioning. Some optimization-based feature partitioning methods are given below:

- *Omel-C-Pso* [21]*:* This approach utilizes particle swarm optimization (PSO) to determine the best feature partitioning. The goal is to minimize the intraclass distance (distance between instances of the same class) and maximize the interclass distance (distance between instances of different classes).

- *MEL Using Multi-Objective PSO* [22]*:* It extends the use of PSO to handle multi-objective optimization in feature partitioning. This approach considers multiple conflicting objectives when dividing features into views.

- *GA-Based FSP* [23]*:* Genetic algorithms (GAs) are employed to optimize feature partitioning. In order to determine the best partitioning, GAs scan the space of potential feature subsets iteratively.

- *Ant System-Based FSP* [24, 25]*:* Drawing inspiration from ant colony optimization, this method optimizes feature partitioning through the exploration of multiple feature subsets. The objective is to select the most informative and non-redundant feature subsets for views.

2.6 Ensemble-based methods

Principles: Ensemble-based feature partitioning methods combine predictions obtained from various FSPs' predictions to create a final prediction. Instead of relying on a single FSP approach, ensemble methods leverage the collective knowledge from multiple FSP methods to enhance classification performance.

Ensemble Multiview FSP (E-FSP) [26]*:* E-FSP combines predictions from various FSP methods to improve classification performance. By aggregating multiple views created using different FSP approaches, it leverages the strengths of these methods to enhance the overall accuracy of the ensemble model.

Each feature partitioning method offers a unique approach to dividing high-dimensional feature spaces into multiple views. The choice of method should consider the specific dataset characteristics, the goals of the machine learning task, and the available computational resources, as each method has its own strengths and weaknesses.

3. Comparative study

Multiview learning, a prominent area in machine learning, often utilizes feature partitioning methods to enhance classification accuracy by creating multiple views of the same dataset. From the state-of-the-art literature's results and analysis [27], we observe that E-FSP and OFSP consistently achieve superior performance across a wide range of datasets. These two methods demonstrate a remarkable ability to enhance classification accuracy in multiview learning contexts (as shown in **Figure 3**). However, it is crucial to note that the choice of a feature partitioning method should not be solely based on accuracy. Every method has a distinct set of advantages and disadvantages that vary based on the particular problem, dataset properties, and available computing power. Selecting the most suitable feature partitioning method should be a thoughtful process, considering all these factors.

3.1 Strengths and weaknesses of each method

Let us provide a more detailed account of the strengths and weaknesses of each feature partitioning method used in the comparative study:

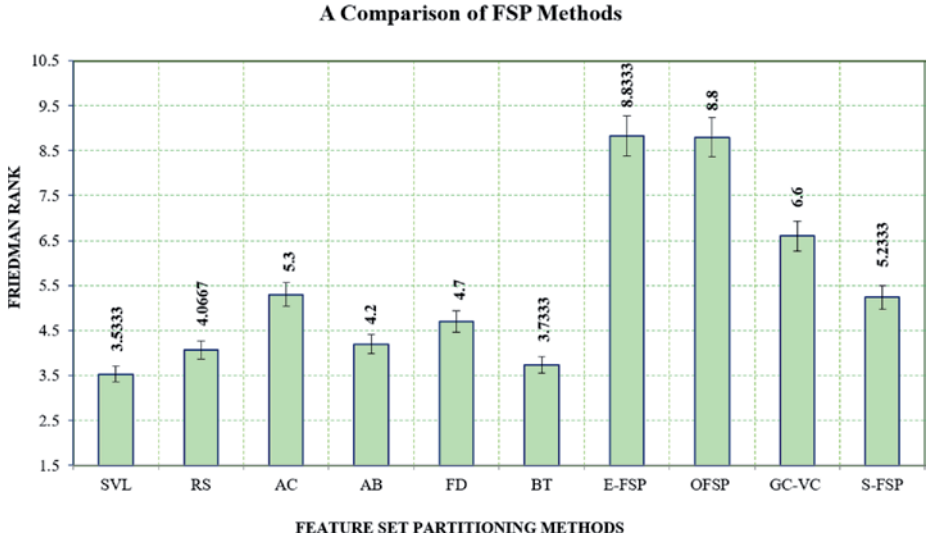

Figure 3.
Comparison of feature partitioning methods.

- *Random Split (RS):* RS is a simple and basic method with straightforward implementation. While it may lead to slight improvements in classification performance, its results are subject to randomness, necessitating multiple experiments for robust conclusions.

- *Attribute Bagging (AB):* AB offers better classification performance than RS, and it is less computationally intensive. However, like RS, it also relies on randomness and may require multiple runs for consistent results.

- *Attribute Clustering (AC):* AC is less time-consuming than RS and AB, as it does not depend on randomness. It partitions features using clustering techniques. Its performance, however, is contingent on the quality of the clustering method employed.

- *Supervised Feature Set Partitioning (S-FSP):* S-FSP is a computationally efficient method because it is based on patterns. It produces more balanced views and performs better than RS and AB, but not as well as approaches focused on optimization.

- *Graph Coloring-Based FSP (GC-VC):* As a performance-based approach, GC-VC's effectiveness depends on determining an ideal threshold, which can be computationally difficult.

- *Bell Triangle-Based FSP (BT):* BT is computationally less expensive and does not require additional parameters, which simplifies its implementation. However, it does not deliver significantly improved performance compared to other pattern-based methods.

- *Ferrers Diagram-Based FSP (FD):* Similar to BT, FD is computationally efficient, creating balanced views. However, it does not outperform optimization-based techniques by a significant margin.

- *Optimal Feature Set Partitioning (OFSP):* OFSP is a performance-based method, and it boasts remarkable performance. It even surpasses genetic algorithm-based methods. However, it is computationally demanding and requires the specification of the number of views in advance.

- *Ensemble Feature Set Partitioning (E-FSP):* An ensemble-based method with the potential for great performance is called E-FSP. Utilizing the combined expertise of many feature partitioning techniques, it improves classification quality. Nevertheless, it comes at the cost of increased computational complexity and may support overfitting, depending on the ensemble methods used.

This comparative study provides valuable insights into the characteristics of different feature partitioning methods, their performance, and the trade-offs associated with each. Researchers and practitioners can use this information to make informed decisions when selecting the most appropriate feature partitioning method for their specific multiview learning tasks. It emphasizes that the choice should align with the problem's specific requirements, computational resources, and dataset properties.

3.2 Issues associated with feature partitioning

Feature partitioning plays a crucial role in multiview learning (MVL), enhancing classification accuracy by creating diverse perspectives of the same dataset. However, like any technique, it comes with its share of challenges and issues. In this section, we explore common challenges and issues associated with feature partitioning in MVL.

- *Selecting the Appropriate Quantity of Views:* Determining the ideal number and sizes of views to be developed is one of the main issues in feature partitioning. The number of views directly impacts the dimensionality and diversity of the feature space. Creating too few views may not capture the intricacies of the data, leading to underfitting, while generating too many views can result in redundancy and overfitting. Deciding on an appropriate number of views often requires a balance that depends on the dataset and the specific problem being addressed.

- *Combining Views:* Once multiple views are created, combining them to form a single consensus decision is another intricate problem. Each view may contribute differently to the overall classification, and conflicts between views can arise. Combining methods, such as majority voting, weighted voting, or more advanced techniques like stacking, are needed to effectively aggregate the views. Selecting the most suitable combination method for a given problem is a nontrivial task.

- *Correlation Between Views:* Feature partitioning methods may inadvertently lead to the creation of views with highly correlated features. Correlation between views can result in redundancy, diminishing the benefits of diversity in MVL. Identifying and addressing feature redundancy across views is essential for the overall effectiveness of the ensemble. Techniques for feature selection or dimensionality reduction may be applied to mitigate this issue.

- *Scalability:* Scalability is a critical concern in feature partitioning, especially when dealing with large datasets. As the number of views and the size of each view increase, the computational complexity of MVL methods also grows. This can lead to extensive memory and processing demands, potentially rendering the application of feature partitioning on large-scale datasets challenging.

- *Data Heterogeneity:* Data heterogeneity can be a significant issue when creating views from distinct feature subsets. Heterogeneity in the data is introduced by the views that are created from several feature subsets, each of which may have unique characteristics and representations. This can make it complex to effectively combine the views, especially when each view is created from a distinct data modality, such as text and images.

- *Computational Complexity:* MVL methods, especially those employing multiple views, can be computationally intensive. Training multiple models simultaneously requires substantial computational resources. Additionally, the complexity of combining views can also increase the overall computational cost. This can be a limiting factor when resources are constrained or when real-time or near-real-time predictions are necessary.

- *Optimal Feature Selection:* It is a non-trivial effort to choose the most informative characteristics from each view and determine which subset of features is optimal for the ensemble. Choosing suboptimal features may lead to a reduction in classification performance, while selecting too many features may increase computational demands and the risk of overfitting. Effective feature selection methods are required to address this challenge.

- *Cross-Domain Generalization:* When applied to diverse datasets or domains with differing feature distributions or class relationships, multiview ensemble models may have trouble generalizing. Adapting multiview ensemble models to perform consistently across diverse datasets and domains is a complex issue.

- *Managing missing information:* It is typical for real-world datasets to have missing values. Handling missing data in different views introduces additional complexity. Handling missing data in the context of feature partitioning requires specialized techniques for accurate and robust predictions.

- *Inconsistency in Labels:* Unequal class distributions among views may result in skewed predictions and impact the ensemble model's overall effectiveness. Handling label imbalance effectively is a key concern in multiview learning.

- *Covariate Transformation:* Covariate modification, in which the distribution of data varies between views, can be introduced by creating views from distinct feature subsets. In an ensemble environment, addressing this change is essential to preserving accurate forecasts.

- *Overfitting:* Overfitting is a potential concern in MEL, particularly when dealing with complex ensemble models or sparse training data sets. A constant problem is striking a balance between model performance and model complexity.

- *Suitable View Weighting:* It is a non-trivial effort to determine the proper weights for each perspective in the ensemble. It can have a major effect on the multiview ensemble's effectiveness as a whole. Effective methods for optimal view weighting are essential.

These challenges and issues associated with feature partitioning in multiview learning underscore the need for continued research and innovation in the field. Addressing these challenges will enable more robust and effective multiview learning methods, ultimately advancing their application in various domains and problem scenarios.

4. Future perspectives

The field of feature partitioning and multiview ensemble learning is a dynamic and evolving area of research. In this section, we explore potential advancements and future directions for the field. Such perspectives seek to address domain-specific issues, increase the effectiveness of classification, diversify as well as enhance the quality of views produced, and facilitate the use of these methods across a range of domains.

- *Integration of Deep Learning Techniques:* One promising avenue for advancing feature partitioning and MEL is the integration of deep learning techniques. Deep neural networks, such as convolutional neural networks (CNNs) for image data or recurrent neural networks (RNNs) for sequential data, have demonstrated exceptional capabilities in feature extraction and representation learning. Integrating deep learning with multiview learning can enhance the creation of views, enabling more automatic and data-driven feature partitioning. Deep learning models can also be employed for view combinations, further improving the accuracy of the ensemble. The exploration of novel deep learning architectures for multiview ensemble learning is an exciting future research direction.

- *Handling Imbalanced and Noisy Data:* Dealing with imbalanced datasets and noisy data is a common challenge in real-world applications. Future advancements in feature partitioning should include specialized techniques to address these issues. Improved methods for creating views that account for class imbalance and noise, as well as novel ensemble strategies, can significantly enhance the robustness of multiview ensemble models. Additionally, advanced approaches for data preprocessing and cleaning within the multiview context will be valuable.

- *Applications in Multi-Modal Data and Real-World Scenarios:* Multiview ensemble learning holds great potential for applications involving multi-modal data, where information is distributed across various data types (e.g., text, images, audio). Future research should focus on developing feature partitioning methods that effectively integrate and leverage diverse data modalities. This will expand the applicability of multiview learning to real-world scenarios, such as healthcare, autonomous driving, and multimedia content analysis.

- *Enhance Interpretability and Explainability:* As multiview ensemble models become more complex, understanding and interpreting the reasoning behind their decisions become essential. Future research should aim to enhance the interpretability and explainability of multiview ensemble models. Developing techniques for visualizing the contributions of individual views, understanding feature importance across views, and providing meaningful explanations for predictions will make multiview ensemble models more transparent and trustworthy.

- *Domain-Specific Challenges:* Different application domains present unique challenges. Future research should consider domain-specific requirements and constraints. For example, in healthcare, ensuring model interpretability and compliance with medical regulations is crucial. In financial services, addressing data privacy and security concerns is paramount. Tailoring feature partitioning and multiview ensemble methods to these domains will unlock their full potential.

- *Evaluation Metrics and Benchmarks:* Developing standardized evaluation metrics and benchmark datasets for feature partitioning and multiview ensemble learning is essential. This will enable fair and consistent comparisons between different methods and facilitate the adoption of best practices. Future research should focus on creating such benchmarks and defining evaluation standards.

- *Scalability and Efficiency:* As the size of datasets continues to grow, the scalability and efficiency of feature partitioning and multiview ensemble methods will be increasingly important. Future advancements should include techniques that optimize memory and processing usage, allowing these methods to be applied to large-scale datasets efficiently.

Overall, the future of feature partitioning and multiview ensemble learning is promising, with opportunities for integration with deep learning, addressing real-world challenges, enhancing interpretability, and expanding into diverse application domains. As researchers continue to innovate and tackle these issues, multiview ensemble learning will become a more accessible and effective tool for a wide range of applications.

5. Conclusion

In this chapter, we have delved into the concept of feature partitioning, a crucial technique in the realm of machine learning and data analysis. Feature partitioning methods aim to divide high-dimensional feature spaces into multiple views, each highlighting different aspects of the data. These views can then be used in multiview ensemble learning to improve classification and prediction tasks. We have explored various feature partitioning methods, including random-based, pattern-based, clustering-based, performance-based, optimization-based, and ensemble-based methods. Our comparative study has shed light on the strengths and weaknesses of different feature partitioning methods. Notably, ensemble-based approaches like E-FSP and optimization-based methods like OFSP have demonstrated exceptional performance, outperforming single-view learning and other feature partitioning methods in most cases. However, they come at the cost of increased computational complexity. The choice of feature partitioning method should be guided by the specific requirements and constraints of the problem at hand. The challenges and issues associated with feature partitioning, such as selecting the right number of views, handling view inconsistency, addressing data heterogeneity, and ensuring scalability, have been discussed. These issues emphasize the need for continued research and innovation in the field to develop more robust and efficient feature partitioning methods. Looking forward, several promising avenues for future research have been highlighted. The integration of deep learning techniques into feature partitioning and multiview ensemble learning is a compelling prospect. Deep neural networks can enhance feature extraction and representation learning, potentially automating and improving the feature partitioning process. Furthermore, the field should address the challenge of imbalanced and noisy data, develop domain-specific solutions, and provide more interpretable and explainable multiview ensemble models. Standardized evaluation metrics and benchmarks are essential to facilitate fair comparisons and promote best practices. Additionally, scalability and efficiency are becoming increasingly important as datasets continue to grow in size and complexity. In conclusion, feature partitioning is a fundamental technique that plays a pivotal role in enhancing the performance and interpretability of machine learning models. Its application spans across various domains, from healthcare to finance, and promises to address real-world challenges. As researchers continue to advance the field and tackle the identified issues, feature partitioning will remain a key tool in the data scientist's toolkit, contributing to more accurate

and robust machine learning solutions. This chapter serves as a stepping stone, guiding researchers toward future prospects and potential areas for research in the ever-evolving landscape of feature partitioning and multiview ensemble learning.

Acknowledgements

The authors acknowledge the usage of QuillBot and ChatGPT for language polishing and professional writing refinement in this manuscript.

Author details

Aditya Kumar* and Jainath Yadav
Department of Computer Science, Central University of South Bihar, Bihar, India

*Address all correspondence to: ark1111adk@gmail.com

References

[1] Zhao J, Xie X, Xu X, Sun S. Multi-view learning overview: Recent progress and new challenges. Information Fusion. 2017;**38**:43-54

[2] Perry R, Mischler G, Guo R, Lee T, Chang A, Koul A, et al. mvlearn: Multiview machine learning in python. The Journal of Machine Learning Research. 2021;**22**(1):4938-4944

[3] Kumar V, Minz S. Multi-view ensemble learning: An optimal feature set partitioning for high-dimensional data classification. Knowledge and Information Systems. 2016;**49**:1-59

[4] Wang Z, Chen S, Gao D. A novel multi-view learning developed from single-view patterns. Pattern Recognition. 2011;**44**(10-11):2395-2413

[5] Gupta A, Khan RU, Singh VK, Tanveer M, Kumar D, Chakraborti A, et al. A novel approach for classification of mental tasks using multiview ensemble learning (mel). Neurocomputing. 2020;**417**:558-584

[6] Alam MT, Kumar V, Kumar A. A multi-view convolutional neural network approach for image data classification. In: 2021 International Conference on Communication Information and Computing Technology (ICCICT), Mumbai, India. IEEE; 2021. pp. 1-6

[7] Ku-Mahamud KR, Sediyono A. A new feature set partitioning method for nearest mean classifier ensembles. In: 4th International Conference on Computing & Informatics, Kuching, Sarawak, Malaysia. 2013

[8] Shi S, Nie F, Wang R, Li X. When multi-view classification meets ensemble learning. Neurocomputing. 2022;**490**:17-29

[9] Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;**20**(8):832-844

[10] Bryll R, Gutierrez-Osuna R, Quek F. Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition. 2003;**36**(6):1291-1302

[11] Guggari S, Kadappa V, Umadevi V. Non-sequential partitioning approaches to decision tree classifier. Future Computing and Informatics Journal. 2018;**3**(2):275-285

[12] Guggari S, Kadappa V, Umadevi V. Frequent itemsets based partitioning approach to decision tree classifier. In: Mining Intelligence and Knowledge Exploration: 7th International Conference, MIKE 2019, Goa, India, December 19-22, 2019, Proceedings 7. Berlin, Heidelberg: Springer; 2020. pp. 286-295

[13] Guggari S, Kadappa V, Umadevi V. Theme-based partitioning approach to decision tree: An extended experimental analysis. In: Emerging Research in Electronics, Computer Science and Technology: Proceedings of International Conference, ICERECT 2018. Singapore: Springer; 2019. pp. 117-127

[14] Kumar V, Minz S. Multi-view ensemble learning: A supervised feature set partitioning for high dimensional data classification. In: Proceedings of the Third International Symposium on Women in Computing and Informatics. New York, NY, United States: Association for Computing Machinery; 2015. pp. 31-37

[15] Kumar A, Kumar V, Kumari S. A graph coloring based framework

for views construction in multi-view ensemble learning. In: 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), Jalandhar, India. IEEE; 2021. pp. 84-89

[16] Kumari S, Kumar V, Kumar A. Effectiveness analysis of distance measures for graph coloring based view-construction approach in multiview ensemble learning. In: Distributed Computing and Optimization Techniques: Select Proceedings of ICDCOT 2021. Singapore: Springer; 2022. pp. 411-424

[17] Janusz A, Slezak D. Rough set methods for attribute clustering and selection. Applied Artificial Intelligence. 2014;**28**(3):220-242

[18] Taheri K, Moradi H, Tavassolipour M. Collaboration graph for feature set partitioning in data classification. Expert Systems with Applications. 2023;**213**:118988

[19] Zheng L, Chao F, Parthal MN, Zhang D, Shen Q. Feature grouping and selection: A graph-based approach. Information Sciences. 2021;**546**:1256-1272

[20] Saini M, Verma S, Sharan A. Multi-view ensemble learning using rough set based feature ranking for opinion spam detection. In: Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2017, Volume 1. Singapore: Springer; 2019. pp. 3-12

[21] Kumar V, Minz S. An optimal multi-view ensemble learning for high dimensional data classification using constrained particle swarm optimization. In: Information, Communication and Computing Technology: Second International Conference, ICICCT 2017, New Delhi, India, 13 May 2017, Revised Selected Papers 2. Singapore: Springer; 2017. pp. 363-378

[22] Kumar V, Aydav PSS, Minz S. Multi-view ensemble learning using multi-objective particle swarm optimization for high dimensional data classification. Journal of King Saud University-Computer and Information Sciences. 2022;**34**(10):8523-8537

[23] Rokach L. Genetic algorithm-based feature set partitioning for classification problems. Pattern Recognition. 2008;**41**(5):1676-1700

[24] Husin A. Ant system-based feature set partitioning algorithm for classifier ensemble construction. International Journal of Soft Computing. 2016;**11**(3):176-184

[25] Mahamud KRK, et al. Ant system-based feature set partitioning algorithm for k-nn and lda ensembles construction. In: 5th International Conference on Computing and Informatics 2015, Istanbul, Turkey. 11-13 August 2015

[26] Singh R, Kumar V. Ensemble multi-view feature set partitioning method for effective multi-view learning. Available from: SSRN 4259844 [Pre-print]

[27] Kumar A, Yadav J. A review of feature set partitioning methods for multi-view ensemble learning. Information Fusion. 2023;**100**:101959

Edited by Jainath Yadav

This book discusses the recent advances and applications of association rule mining techniques in data mining tasks. Association Rule Mining is an emerging and interdisciplinary field of research in data mining. It draws on ideas from various fields to extract association rules or patterns from large databases. Researchers and companies frequently use the concepts of association rule and data mining for prediction/classification tasks. In today's digital era, vast amounts of data are being generated and stored in databases. Therefore, there is a need to analyze the data and identify important association rules and patterns from this vast dataset. *Recent Advances in Association Rule Mining and Data Mining* is an important step in this direction. The primary aim of the book is to provide the reader with a comprehensive overview of recent advances and developments in this field. The book is a collection of reviewed book chapters written by researchers, academicians, and other research scientists working in the areas of Association Rule Mining and Data Mining.

Andries Engelbrecht, Artificial Intelligence Series Editor

ISSN 2633-1403
ISBN 978-0-85466-638-6
Published in London, UK

IntechOpen