A NOVEL ALGORITHM FOR IMPROVED TRACTABILITY AND EXPLAINABILITY OF INFERENCES ON HIGH-DIMENSIONAL DATA USING CONDITIONED ATTENTION MECHANISM

ASHUZA KUDERHA RIFIN
(22PCG02636)
B.Sc Computer Science, Université Catholique de Bukavu, Bukavu, DR Congo

AUGUST, 2024

A NOVEL ALGORITHM FOR IMPROVED TRACTABILITY AND EXPLAINABILITY OF INFERENCES ON HIGH-DIMENSIONAL DATA USING CONDITIONED ATTENTION MECHANISM

BY

ASHUZA KUDERHA RIFIN
(22PCG02636)
B.Sc. Computer Science, Université Catholique de Bukavu, Bukavu, DR Congo

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MASTER OF SCIENCE (M.Sc.) DEGREE IN COMPUTER SCIENCE IN THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COVENANT UNIVERSITY, OTA, OGUN STATE

AUGUST, 2024

ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfillment of the requirements for the award of the degree of Master of Science in Computer Science in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria.

Miss Adefunke F. Oyinloye
(Secretary, School of Postgraduate Studies) **Signature and Date**

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies) **Signature and Date**

DECLARATION

I, ASHUZA KUDERHA Rifin (22PCG02636) declare that the research entitled "**A NOVEL ALGORITHM FOR IMPROVED TRACTABILITY AND EXPLAINABILITY OF INFERENCES ON HIGH-DIMENSIONAL DATA USING CONDITIONED ATTENTION MECHANISM**" was carried out by me under the supervision of Dr. Iheanetu Olamma of the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. I attest that this dissertation has not been submitted either wholly or partially for the award of any degree elsewhere. All data and scholarly information used in this dissertation are duly acknowledged.

ASHUZA KUDERHA Rifin

Signature and Date

CERTIFICATION

This is to certify that this dissertation titled "**A NOVEL ALGORITHM FOR IMPROVED TRACTABILITY AND EXPLAINABILITY OF INFERENCES ON HIGH-DIMENSIONAL DATA USING CONDITIONED ATTENTION MECHANISM**" is original research carried out by **ASHUZA KUDERHA Rifin (22PCG02636)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria under the supervision of **Dr. Iheanetu Olamma**. We have examined and founded this work acceptable as part of the requirements for the award of Master of Science (M.Sc.) in Computer Science.

Dr. Iheanetu Olamma
(Supervisor) **Signature and Date**

Prof. Olufunke Oladipupo
(Head of Department) **Signature and Date**

Prof. Afolashade O. Kuyoro
(External Examiner) **Signature and Date**

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies) **Signature and Date**

DEDICATION

I dedicate this work to all the healthcare workers who dedicated their lives to save the lives others. May this work help to measure their contributions.

ACKNOWLEDGEMENTS

I would like to express my gratitude to God who made this journey in Nigeria possible. To my family, I say thank you for the support.

To all my lectures, I say thank you for the training. Especially, I thank my Supervisor, Dr. Iheanetu Olamma, for her guidance and support.

I am also grateful for the support from CApIC-ACE to help me pursue my dreams. I appreciate the work of Prof. Emeka E.J.Iweala, Prof. E. Adetiba, Prof. Ezekiel Adebiyi, Prof. Olubanke Ogulana and Ms. Helen Jevwegaga.

To all my fellows, I say thank you for the support during this Master's Program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The availability of a large amount of data that can be mined is a challenge for data mining techniques that cannot keep up with the exponentially increasing quantity of data. Most of the classical techniques have been reported inefficient to handle large amount of data. The aim of this work is to develop a novel attention mechanism for high-dimensional tabular data understanding for knowledge representation that can enhance tractability of decision rule learning. The conditioning theory was mathematically formalized as an attention mechanism that models an unconditioned response (or reflex response) to unconditioned stimulus by pairing conditioned stimulus with unconditioned stimulus. The formalization of the conditioning theory built was translated into an algorithm presented as conditioned attention mechanism and applied on input tabular dataset producing vector embeddings of attribute-value pairs. The learned representation of attribute-value pairs as vector embeddings are then exploited to build decision rules using a developed novel Coupling-Sorting-Branching technique (CoSoBra), formalized as energy-based model optimization. A Python-based compression algorithm was developed to reduce the dimension of the resulting very large vector embeddings while maintaining the quality of the attribute-value pairs representations. The developed novel Conditioned Attention High Dimensional Rule Learning (CAHDRL) algorithm, which captures both CoSoBra algorithm and Condition Attention mechanism was benchmarked against existing approaches for decision rule learning on ten datasets from six different domains. The efficacy of the CAHDRL approach was evaluated with respect to the effectiveness of the extracted decision rules and its efficiency during data processing. An ablation study was carried out to assess the efficiency CAHDRL. Results of the evaluation on the 8 first datasets showed that CADHRL had higher accuracy of 0.85 in average against 0.8175 in average for the EAV-60 model, equal length of decision rule of 2.125 in average against 2.125 in average for the reduct-alpha-0.1 model and a better support of 392.5263 in average against 350.2663 in average for the reduct-alpha-0.1. t-distributed Stochastic Neighborhood Embedding (t-SNE) was used to visualize the embeddings on a 2D manifold. The results support literature findings on Mushroom dataset that “odor=creosote” pair is a good descriptor of the poisonous mushroom. The second visualization results using “Adult” dataset showed that USA natives are more likely to earn $50,000.00 and above. Wilcoxon test of significance of efficacy results for CAHDRL validated its relatively superior performance against existing approaches, yielding values of p-values inferior to 0.05 for accuracy (3 of the seven comparisons), length of decision rules (6 of the seven comparisons) and support (4 of the five comparisons) respectively. CAHDRL can effectively find embedding vectors of attribute-value pairs and efficiently extract decision rules from diverse datasets for classification tasks, with a higher tractability and inherent explainability. CAHDRL also demonstrates a great scalability feature that allows for the efficient processing very high dimensional datasets by the parallelizable nature of CAHDRL.

Keywords: attention mechanism, decision rule, explainability, tractability, machine learning