UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Development of Robust and Scalable Hyperbox based Machine Learning Algorithms

by

Thanh Tung KHUAT

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2021

Certificate of Original Authorship

I, Thanh Tung KHUAT, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Date: 01 July 2021

Abstract

Together with the rapid development of digital information and the increase in amount of data, machine learning (ML) algorithms have been developed and evolved constantly to discover new information and knowledge from different data sources. The use of hyperbox fuzzy sets as fundamental representational and building blocks in learning algorithms forms an important branch of ML. Hyperbox-based algorithms have a huge potential for high scalability and incremental adaptation to applications working in the dynamically changing environments. Additionally, learning algorithms based on hyperbox representations can form interpretable models, which are highly desirable for areas with the requirement of safety and trust. This study aims to develop and expand robust, scalable, and transparent learning algorithms for hyperbox-based classification models with a specific focus on a general fuzzy min-max neural network (GFMMNN).

First of all, a comprehensive survey on hyperbox-based machine learning models together with empirical assessments of the GFMMNN on pattern classification problems were conducted. Next, a new online learning algorithm was proposed for the GFMMNN and improved the robustness of the whole family of GFMMNN learning algorithms to work effectively with mixed-attribute data by introducing a new learning mechanism for categorical features. In terms of scalability, the main steps of the learning algorithms were reformulated so they can be effectively executed on graphics processing units using matrix operations, simultaneously proposing mathematical lemmas to reduce the redundancies of hyperbox candidates in the learning process. This thesis also proposed a novel method to enhance the transparency of classifiers while maintaining a good classification performance by using hierarchical granular representations from hyperbox fuzzy sets. The last contribution was a simple but powerful ensemble model built from many individual hyperbox-based classifiers trained on random subsets of both sample and feature spaces. Extensive empirical analyses indicated that the proposed solutions are highly competitive with other evaluated learning algorithms.

To my loved ones

Acknowledgments

The Ph.D. study is a long journey, and this thesis would not have been possible without the support and encouragement of my supervisors, friends, and relatives during my Ph.D. research journey. Maybe the hardest part when I write this thesis is how to express my sincere gratitude to them.

First and foremost, I would especially like to express my sincerest gratitude to my principal supervisor, Professor Bogdan Gabrys for his continuous support, motivation, enthusiasm, inspiration to my Ph.D. study, together with his invaluable advice and discussion during our weekly meetings. His guidance is likely to positively affect me throughout my future career path. He will definitely be the first person to whom I seek advice for the difficult research problems that I have to deal with in the future. My Ph.D. research has faced a difficult time because of the spread of the COVID-19 pandemic. However, under his supervision and instructions, I has been quick to change research plans and adapt to the new ways of working using remote working tools. As a result, my research progress has not been negatively impacted by the pandemic. Those are precious experiences that will never be forgotten.

I would also like to thank my co-supervisors, Distinguished Professor Fang Chen and Dr. Dymitr Ruta for their valuable comments and discussion on the research manuscripts that I have written for my Ph.D. projects. My thanks also go to my Ph.D. assessment panel, Professor Paul Kennedy and Associate Professor Wei Liu, for their constructive feedback. Special thanks also to Dr. Hanh Le for encouraging me to pursuit Ph.D. research. I wish to thank UTS for awarding me scholarships so that I can conduct my Ph.D. project.

A large part of this thesis comes from publications which have been peer-reviewed by anonymous reviewers. Therefore, I also wish to send a special thank to them who made valuable suggestions to enhance the quality of my research papers.

Additionally, I would like to thank all my best friends who have accompanied me throughout my Ph.D. journey: Tien-Dung Nguyen and his family, Thong Do, Thac Do, Cong Nguyen, Tung Huynh, An Le, Dung Duong, Xuan Yang, Joakim Skarding, Sunny Verma, and many others. Happy conversations and all the fun we have had during the past three years will be unforgotten memories in my life.

Finally, I would like to express my love and boundless gratitude to my parents and younger sister for their unconditional support and sacrifice. They have supported me both financially and emotionally so that I can completely concentrate on my Ph.D. research.

Thanh Tung Khuat

Sydney, Australia, 2021.

List of Publications

Published Papers

1. **Thanh Tung Khuat**, and Bogdan Gabrys, "Random hyperboxes," *IEEE Transactions on Neural Networks and Learning Systems* (Early Access). (*Chapter 8*)
2. **Thanh Tung Khuat**, and Bogdan Gabrys, "An in-depth comparison of methods handling mixed-attribute data for general fuzzy min-max neural network," *Neurocomputing*, vol. 464, pp. 175-202, 2021. (*Chapter 5*)
3. **Thanh Tung Khuat**, Fang Chen, and Bogdan Gabrys, "An effective multi-resolution hierarchical granular representation based classifier using general fuzzy min-max neural network," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 2, pp. 427- 441, 2021 (**IEEE CIM Publication Spotlight paper**). (*Chapter 7*)
4. **Thanh Tung Khuat**, and Bogdan Gabrys, "Accelerated learning algorithms of general fuzzy min-max neural network using a novel hyperbox selection rule," *Information Sciences*, vol. 547, pp. 887-909, 2021. (*Chapter 6*)
5. **Thanh Tung Khuat**, Dymitr Ruta, and Bogdan Gabrys, "Hyperbox based machine learning algorithms: A comprehensive survey," *Soft Computing*, vol. 25, pp. 1325–1363, 2021. (*Chapter 2*)
6. **Thanh Tung Khuat**, and Bogdan Gabrys, "A comparative study of general fuzzy min-max neural networks for pattern classification problems," *Neurocomputing*, vol. 386, pp. 110-125, 2020. (*Chapter 3*)
7. **Thanh Tung Khuat**, Fang Chen, and Bogdan Gabrys, "An improved online learning algorithm for general fuzzy min-max neural network," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1-9,

2020 (**IEEE CIS Outstanding Student Paper Conference Registration Grant**). (*Chapter 4*)

8. **Thanh Tung Khuat**, and Bogdan Gabrys, "Accelerated training algorithms of general fuzzy min-max neural network using gpu for very high dimensional data," in *Proceedings of the 26th International Conference on Neural Information Processing (ICONIP)*, pp. 583-595, 2019 (**Best paper award finalists**). (*Chapter 6*)

Under Reviewed Papers

9 **Thanh Tung Khuat**, and Bogdan Gabrys, "An online learning algorithm for a neuro-fuzzy classifier with mixed-attribute data," Submitted to *IEEE Transactions on Fuzzy Systems* (Revised and Resubmit). (*Chapter 5*)

Contents

List of Figures

List of Tables

List of Abbreviations

2lv-CSWCE	Two-level classification system with testing in dynamically changing environment
2-pHC	Two-phase hyperbox classifier
ACO	Ant colony optimization
Adaboost	Adaptive boosting
AGGLO-SM	Agglomerative learning algorithm using the full similarity matrix
AGGLO-2	Agglomerative algorithm version 2
ANNs	Artificial Neural Networks
ARC	Adaptive resolution classifier
ART	Adaptive Resonance Theory
AUC	Area under the curve
CACL	Concept adapting contraction less
CBA	Class balanced accuracy
CD	Critical difference
CLFMNN	Contraction-less fuzzy min-max neural network
CNs	Compensatory Neurons
DCFMN	Data core based fuzzy min–max neural network
DPS	Density-preserving sampling
ECTs	Ensemble model of clustering trees
EFMNN	Enhanced fuzzy min-max neural network
EFMNN-ACO	Enhanced fuzzy min–max neural network with ant colony optimization
EFMMDT	Evolving fuzzy min–max decision tree
EFMNN-II	Enhanced fuzzy min-max neural network with K-nearest hyperbox expansion rule and pruning
EFMNN-UL	Evolved fuzzy min-max neural network for unknown labeled data
EIOL-GFMM	Extended improved online learning algorithm for general fuzzy min-max neural network
EMILP	A enhanced version of mixed integer linear programming model-based hyperbox classifier
Esb-GFMMNN	Ensemble of neuro-fuzzy classifiers
Ens-IOL-GFMM(DL)	Ensemble of GFMMNNs using the IOL-GFMM algorithm at the decision level
Ens-IOL-GFMM(ML)	Ensemble of GFMMNNs using the IOL-GFMM algorithm at the model level
FMCN	Fuzzy min-max neural network with compensatory neurons
FMM-CART	Offline and online fuzzy min–max neural network and classification and regression trees
FMM-CT	Fuzzy min-max clustering neural network with the clustering tree
FMM-GA	Fuzzy min-max neural network with genetic algorithms
FMMDT	Fuzzy min–max neural network based decision tree
FMM-PSO	Fuzzy min-max neural network with the particle swarm optimization
FMM-ECT	Fuzzy Min–Max neural network with an ensemble of clustering trees
FMNWSM	Fuzzy min-max neural network with symmetric margin
FMNN	Fuzzy min-max neural network for classification
FMNN-clu	Fuzzy min-max neural network for clustering
FPGAs	Field-programmable gate arrays

GA	Genetic algorithms
GFMM	General fuzzy min-max
GFMMNN	General fuzzy min-max neural network
GFMMNN-CD1	General fuzzy min–max neural networks for categorical data
GFMMNN-CD2	Enhanced general fuzzy min–max neural networks for categorical data
GNB	Gaussian Naive Bayes
GPU	Graphics Processing Unit
GRFMN	General reflex fuzzy min-max neural
HACO	Hyperbox based clustering with ant colony optimization
HACO2	Hyperbox classifier with ant colony optimization
HFC	Hyperbox fuzzy classifier
HNN	Hyperbox neural network algorithm
IEFCN	Inclusion/exclusion fuzzy hyperbox classification network
IGs	Information granules
IOL-GFMM	Improved online learning algorithm for general fuzzy min-max neural network
KNEFMNN	Enhanced fuzzy min-max neural network with K-nearest hyperbox expansion rule
KNN	K-nearest neighbors
LightGBM	Light Gradient Boosting Machines
LOO	Leave-One-Out
MILP	Mixed integer linear programming model-based hyperbox classifier
MDCFMN	Modified data-core-based fuzzy min–max neural network
MEFMNN	Modified-enhanced fuzzy min–max neural network
MFMC	Modified fuzzy min-max neural network for clustering
MFMNN	Modified fuzzy min-max neural network for two-stage pattern classification
EFMNNC	Enhanced fuzzy min-max neural network for clustering
MFMNN-GA	Modified fuzzy min-max neural network with genetic algorithms
MLF	Multi-level fuzzy min-max neural network
MMM-BL	Modified fuzzy min-max neural network with a new batch learning algorithm
MRHGRC	Multi-resolution hierarchical granular representations based classifier
Onln-GFMM	Original online learning algorithm for general fuzzy min-max neural network
PARC	Pruning Adaptive resolution classifier
RBF	Radial Basis function
ReFMN	Reflex Fuzzy Min Max Neural Network
ReFMN-FN	Reflex Fuzzy Min Max Neural Network with floating neurons
RFMNN	Refined Fuzzy Min-Max Neural Network
RH	Random Hyperboxes
RMILP	Refined mixed integer linear programming model-based hyperbox classifier
SAS	Smart adaptive systems
SFMN	Stochastic fuzzy min-max neural network
SS-FMM	Semi-supervised classification method based on fuzzy min–max neural network
SVM	Support Vector Machines
TDFMM	Top-down fuzzy min-max
TDFMMR	Top down fuzzy min-max regressor
TEH-GFMMNN	Tree ensemble hyperboxes via general fuzzy min-max neural network
XGBoost	Extreme Gradient Boosting
WFMM	Weighted fuzzy min-max neural network

List of Notations

If there is no specific definition in each section, the following are the default meanings of notations used in this thesis.

Symbol	**Meaning**
B_i	The *i-th* hyperbox in the list of hyperboxes
$X = [X^l, X^u]$	An input sample in the form of lower and upper bounds
$\mathbf{X}$	A sample space contains all input samples
$\mathcal{T}_N = \{(X_i, c_i)\}_{i=1}^N$	A training data set with N samples
c_i	Class label of sample X_i
c_X	Class label of sample X
$\mathcal{C}$	A set of categorical variables denoting classes to which the observations fall into
n	Number of features of input samples. If a input sample contains both continuous and categorical features, then n denotes the number of continuous features
r	Number of categorical features
$d(A, B)$	Euclidean distance between two vectors A and B
$X^l = (x_1^l, \ldots, x_n^l)$	Lower bound of an input sample in n-dimensional space
$X^u = (x_1^u, \ldots, x_n^u)$	Upper bound of an input sample in n-dimensional space
$X^d = (x_1^d, \ldots, x_r^d)$	A vector contains r categorical attributes for an input sample
$V_i = (v_{i1}, \ldots, v_{in})$	A vector represents the minimum point of the hyperbox B_i in n-dimensional space
$W_i = (w_{i1}, \ldots, w_{in})$	A vector represents the maximum point of the hyperbox B_i in n-dimensional space
G_i	Sample centroid of the hyperbox B_i
n_i	The number of current samples included in the hyperbox B_i
$b_i(X, V_i, W_i)$	Membership value between the input sample X and the hyperbox B_i
$\mathcal{V}$	A matrix contains all minimum points V_i of all hyperboxes B_i generated in the learning process
$\mathcal{W}$	A matrix contains all minimum points W_i of all hyperboxes B_i generated in the learning process
$\mathcal{L}$	A vector contains all class labels for all hyperboxes B_i generated in the learning process
θ	Maximum hyperbox size
Θ	A list of maximum hyperbox sizes
σ	Minimum similarity threshold so that two hyperboxes B_i and B_k can be aggregated
γ	Sensitivity parameter to control the decreasing speed of membership values
$H_j(B_i)$	The current entropy of hyperbox B_i for the *j-th* categorical feature
Ω_{ij}	A set of categorical values on the *j-th* categorical attribute of the hyperbox B_i
s_{ik}	Middle gap similarity measure between two hyperboxes B_i and B_k
$\widetilde{s}_{ik}$	shortest gap similarity measure between two hyperboxes B_i and B_k
$\widehat{s}_{ik}$	longest gap similarity measure between two hyperboxes B_i and B_k

Symbol	Meaning
Φ	A randomizing vector
Φ_i	An independent and identically distributed random vector
M	Number of base learners in an ensemble model
$\mathbb{1}(\cdot)$	Indicator function
$\mathbb{E}$	Expectation
r_s	Sampling rate to find the number of training samples for base learners in an ensemble model
m_f	The maximum number of used features for each base learner in an ensemble model
$\mathcal{E}^*$	Upper bound of the generalization error
$\overline{\rho}$	Average correlation between base learners in an ensemble model
$\overline{K}$	The average number of hyperbox candidates with the same class as the input sample
$\overline{R}$	The average number of hyperboxes representing classes different from the class of the input pattern in each iteration