Faculty of Computer Science and Information Technology

Information Theoretic-based Feature Selection for Machine Learning

Muhammad Aliyu Sulaiman

Doctor of Philosophy
2018

UNIVERSITI MALAYSIA SARAWAK

Grade: ____________

Please tick (√)

Final Year Project Report ☐

Masters ☐

PhD ☑

DECLARATION OF ORIGINAL WORK

This declaration is made on the ....01.......day of November 2018.

Student's Declaration:

I ---------- Muhammad Aliyu Suleiman (14010100) FCSIT ----------

(PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled, Information Theoretic-based Feature Selection for Machine Learning ----- is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

01-11-2018
Date submitted

Muhammad Aliyu Suleiman (14010100)
Name of the student (Matric No.)

Supervisor's Declaration:

I---- JANE LABADIN ---- (SUPERVISOR'S NAME) hereby certifies that the work entitled, INFORMATION THEORETIC-BASED FEATURE SELECTION FOR MACHINE LEARNING (TITLE) was prepared by the above named student, and was submitted to the "FACULTY" as a * ~~partial~~/full fulfillment for the conferment of ---- PhD in Computer Science ---- (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: JANE LABADIN
(Name of the supervisor)

Dr Jane Labadin
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak

Date: 5/11/18

I declare this Project/Thesis is classified as (Please tick (√)):

☐ CONFIDENTIAL (Contains confidential information under the Official Secret Act 1972)*
☐ RESTRICTED (Contains restricted information as specified by the organisation where research was done)*
☑ OPEN ACCESS

Validation of Project/Thesis

I therefore duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies for the purpose of academic and research only and not for other purpose.
- The Centre for Academic Information Services has the lawful right to digitise the content to for the Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic exchange between Higher Learning Institute.
- No dispute or any claim shall arise from the student itself neither third party on this Project/Thesis once it becomes sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student except with UNIMAS permission.

Student's signature ________________ Supervisor's signature: ________________

(Date) (Date)

Current Address:
Deanship of Graduate Studies, King Fahd University of Petroleum and Minerals Dhahran, 31261, Saudi Arabia

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the period and reasons of confidentiality and restriction.

[The instrument was duly prepared by The Centre for Academic Information Services]

Information Theoretic-based Feature Selection for Machine Learning

Muhammad Aliyu Sulaiman

A thesis submitted

In fulfilment of the requirements for the degree of Doctor of Philosophy

(Computer Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2018

DECLARATION

The work described in this PhD thesis, entitled **"Information Theoretic-based Feature Selection for Machine Learning"** is to the best of the author's knowledge that of the author except where due reference is made.

The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

(Student's signature)
ID: 14010100
Muhammad Aliyu Sulaiman

February 2018
Date

ACKNOWLEDGEMENT

First and foremost, I would like to give thanks to almighty Allah (He who taught man the use of the pen) for giving me the ability and strength to see to the completion of this thesis work.

Acknowledgment is due to Universiti Malaysia Sarawak for giving me the opportunity to study and making my dreams come true.

My special and sincere gratitude goes to my Ph.D. thesis supervisor, Assoc. Prof. Dr. Jane Labadin for her support, guidance and encouragement throughout this work. I remained grateful for her critical review and superior suggestions in the course of this thesis work.

I am thankful to the Dean, Centre for Graduate Studies, Prof. Dr. Fasihuddin Badruddin Ahmad, the Dean, Faculty of Computer Science and Information Technology, Assoc. Prof. Dr. Johari Abdullah, Head of the Department of Computational Science & Mathematics, Dr. Chiew Kang Leng and other academic & non-academic staffs of the faculty for their cooperation and supports.

My acknowledgment goes to the members of my extended family; my parent, my wife and kids, my uncles, my aunts, brothers, sisters, friends and family friends for their unshakable caring feelings towards my studies and life in general.

I remain indebted to friends I met in UNIMAS whom time and space will not allow me to enumerate all of them here. However, I am thankful to Mr. Adam Mani Yangora and Dr. Shakeel for their enormous support throughout my time of study in UNIMAS.

ABSTRACT

Three major factors that determine the performance of a machine learning are the choice of a representative set of features, choosing a suitable machine learning algorithm and the right selection of the training parameters for a specified machine learning algorithm. This thesis tackles the problem of feature selection for supervised machine learning prediction tasks through dependency information. The feature evaluation strategy is formulated based on mutual information (MI) to handles both classification and regression supervised learning tasks and the search strategy is a modified greedy forward strategy designed to manage redundancy between features and avoiding features that are irrelevant to the predicting output. The problem with many existing feature selections that evaluate features based on mutual information is that they are designed to handles classification tasks only. And the few existing ones that can work for regression tasks were recently found to underestimate mutual information between two strongly dependent variables. In addition to these problems, the search strategy which is usually a heuristic greedy method used with many existing feature selections, lacks scientifically sound stopping criterion and the forward greedy procedure despite its advantages over the backward procedure is found to reveal suboptimal. Thus, this thesis has developed and evaluated a filter based Information Theoretic-based Feature Selection (IFS) for machine learning. Various experiments were carried out to assess and test components of IFS algorithm. The first test was designed to evaluate the formulated IFS Selection Criterion Strategy (MI estimator) by comparing it with six different MI estimator benchmarks. The second test evaluates IFS in a controlled study using simulated datasets. Moreover, the third test used ten natural domain datasets obtained from UCI Repository, in about fifteen different experiments, using three to four different Machine Learning

Algorithms for performance evaluation. Also, additional experiments to compare the relative performance of the IFS with five related feature selection algorithms were carried out using natural domain datasets. Besides, this thesis developed a hybrid filter method to enhance the performance of the IFS. IFS served as filter together with an Ant Colony Optimization System (ACO) as a metaheuristic form the hybrid system. In these extended IFS method, feature selection method was defined and presented as a 0-1 Knapsack Problem (MKP). Thus, this thesis precisely developed and evaluated IFS_BACS (Binary Ant Colony System) hybrid method. Further experiments were carried out using the natural domain datasets and comparison were made between IFS and hybrid IFS_BACS methods. In most of the cases, experimental results of IFS and its extended IFS_BACS hybrid method significantly reduced features and produce competitive performance accuracy when compared to the results of the full feature set before applying the IFS or IFS_BACS method. And comparing the IFS with its extended version, the extended version (IFS_BACS) seems to be more promising in selecting optimal feature subset from large datasets.

Keywords: Feature Selection, Information Theory, Mutual Information, Entropy, Density Estimation, Optimization, Machine Learning Algorithms, Supervised Learning, Modeling, Ant Colony System.

Maklumat Berasaskan Teori Pemilihan Ciri untuk Pembelajaran Mesin

ABSTRAK

Tiga faktor utama yang menentukan prestasi pembelajaran mesin adalah pilihan set wakil ciri-ciri, memilih algoritma pembelajaran mesin yang sesuai dan pilihan yang tepat parameter latihan bagi algoritma pembelajaran mesin yang ditetapkan. Tesis ini menangani masalah pemilihan ciri untuk tugas-tugas ramalan pembelajaran mesin diselia melalui maklumat kebergantungan. Strategi penilaian ciri dirumuskan berdasarkan maklumat bersama (MI) untuk mengendalikan kedua-dua klasifikasi dan regresi di bawah seliaan tugas-tugas pembelajaran dan strategi carian adalah satu strategi ke hadapan tamak diubahsuai direka untuk menguruskan lebihan antara ciri-ciri dan ciri-ciri mengelakkan yang tidak relevan dengan output meramalkan. Oleh itu, maklumat penapis berdasarkan pemilihan ciri berdasarkan teori-(IFS) untuk pembelajaran mesin telah dibangunkan dan dinilai. Pelbagai eksperimen telah dijalankan untuk menilai dan komponen ujian algoritma IFS. Ujian pertama telah direka untuk menilai IFS pemilihan strategi kriteria yang digubal (MI penganggar) dengan membandingkannya dengan enam garis dasar penganggar MI berbeza. Ujian kedua yang telah dilakukan untuk menilai IFS dalam kajian yang dikawal menggunakan set data yang tiruan. Ujian ketiga menggunakan sepuluh dataset domain semula jadi yang diperolehi daripada UCI repositori, kira-kira lima belas eksperimen yang berbeza, menggunakan 3-4 algoritma pembelajaran mesin yang berbeza untuk penilaian prestasi. Dalam usaha untuk meningkatkan prestasi IFS, kaedah penapis-wrapper hibrid telah dibangunkan. IFS telah digunakan sebagai penapis bersama-sama dengan semut sistem koloni pengoptimuman metaheuristic (ACO) sebagai pembungkus. Dalam kaedah IFS lanjutan ini, kaedah pemilihan ciri ditakrifkan dan dibentangkan sebagai 0-1 masalah buntil multidimensi (MKP). Oleh itu, kaedah hibrid IFS_BACS (binari sistem koloni semut) telah *dibangunkan dan dinilai. Eksperimen selanjutnya telah dijalankan dengan menggunakan set data domain semula jadi dan perbandingan antara IFS dan IFS_BACS kaedah hibrid telah dibuat. Dalam kebanyakan kes, keputusan eksperimen kaedah hibrid IFS dan IFS_BACS dilanjutkan berkurangan ciri-ciri dan menghasilkan ketepatan prestasi kompetitif berbanding keputusan ciri penuh ditetapkan sebelum menggunakan kaedah IFS atau IFS_BACS.*

Kata kunci: *Pemilihan Ciri, Teori Maklumat, Maklumat Bersama, Entropi, Ketumpatan Anggaran, Pengoptimuman, Algoritma Pembelajaran Mesin, Pembelajaran Terselia, Pemodelan, Sistem Koloni Semut.*

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As the quest for computers to be sufficiently intelligent like human (McCarthy & Hayes, 1987; Harnad, 2008) advances in the recent time, the exciting field of Machine Learning emerges (Langely, 2011) to provides a new capability for computers. Applications such as credit card fraud detection, spam detection, speech understanding, product recommendation, medical diagnosis, stock trading, handwritten recognition, natural language processing and other works in computer vision, all of which proved to be very difficult or nearly impossible to program by hands are now possible with the recent development of machine learning technologies. At the heart of the machine learning technologies is the process of learning from a feature set**,** selected from an available dataset. However, the mostly available dataset in various application areas contain features or attributes collected for some reasons other than mining the data, consequently, they contain replicate or irrelevant features to the predicting tasks (Fayyad et al., 1996), of great importance to the learning machine, is feature selection. Feature selection is an optimization method used to select optimal feature subset from the full or whole feature set. Selected optimal feature subset is expected to retain the relevant dependency information in the full feature set as a key focus of a learning algorithm to the predictable tasks. Thus, the hypothesis investigated in this thesis is that feature selection for supervised learning tasks can be achieved based on the dependency information between input features and the target attribute. Also, that such feature selection technique will be useful to wide range of common machine learning algorithms. A technique for Information Theoretic-based Feature Selection (IFS) is developed and evaluated using some popular machine learning algorithms and empirical methods. The selection criterion function is derived based on the popular information theory introduced by Shannon (Shannon, 1948). It is simple to execute and can eliminate unnecessary features. In many cases, it improves the performance of learning algorithms and produces results that are comparable with state of the art feature selection methods from the available literature.

1.1 Background of the Study

Machine learning is an exciting field of study that evolved out of work in artificial intelligence (Langely, 2011). It brings together learning theories from computer science and statistics (Jordan, 2014) and primarily concerns with prediction tasks, based on known or learned features of a dataset. There are two types of machine learning prediction tasks. The first category is classification task which is a prediction task that classifies target object (Output Attribute) into one of the ordinal, nominal or categorical pre-labeled attribute known as class. For example, a machine learning task for medical diagnosis to predict if a patient has cancer or not. The second type of machine learning prediction tasks is the regression task which involves predicting target object as a quantitative or numerical approximation. For the same cancer example, for instance, the prediction task could be to predict the approximate size or spread of cancer in a patient.

Irrespective of the type of tasks, machine learning problems are divided into three broad categories. The first division is the supervised learning; this requires that the dataset should consist of input attributes or features and the corresponding target attribute. The goal here is that the learning system learns model parameters that maps inputs (unseen before) to outputs. The second category is the unsupervised learning, which involves unlabeled data (Dataset consist of only inputs features). The goal of unsupervised learning is to learn a hidden structure or pattern in a dataset. The third category of machine learning task is reinforcement learning, which requires a learning system to learn how to interact with its environment in a dynamic way. The goal is that the learning system learns how to map situation to action like in the case of learning to play a game by playing against an opponent (Bishop, 2006).

State of the art machine learning algorithms, such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian Network, Decision Tree Learning, Clustering, Evolutionary Algorithms, Reinforcement Learning and so forth have been explored and modeled to successfully solve a wider range of prediction problems in various areas of human endeavor. The performance of any learning algorithm to successfully predict an unseen data relies heavily on the nature of data attributes used for learning the model parameters. Various studies carried out in the literature indicated that using "Relevant" attributes or features to train a learning model results in compact models that generalized better (Kira & Rendell, 1992; Koller & Sahami, 1995; Guyon & Elisseeff, 2003; Pudil & Somol, 2005; Gadat & Younes, 2007). Feature selection algorithms are used to select feature subset in a preprocessing step. So, the chosen feature subset is then used for learning the model parameters.

Feature Selection techniques will continue to play an important role in machine learning and pattern recognition for many years to come (Guyon & Elisseeff, 2003; Collins & Liu, 2005; Hoi et al., 2012; Wang et al., 2013; Wu et al., 2014). Most especially, now that the world is currently facing challenges of big data analytics. Pending on the learning domain and processing, Feature Selection Algorithms consist of two broad categories namely the batch and online feature selection techniques. Batch feature selection methods fundamentally assume that all or fixed set of candidate features for learning are available in advance. Moreover, a feature selection method is to choose an m feature subset from a pool of n candidates, where m is expected to be less than the n. This choice is made by optimizing a criterion function over all subsets of the m (Collins & Liu, 2005). This category of feature selection is suitable for standard or traditional learning process (Glocer et al., 2005; Wang et al., 2013). The online feature selection methods assume that features arrive in stages, and no new addition of instances to the problem (Glocer et al., 2005). That is, feature flows into a model one after the other dynamically and feature selection performs at the arrival time. This online feature selection differs from standard online training which assumes instances arrive dynamically (Hoi et al., 2012; Wang et al., 2013). This thesis focuses on batch feature selection for supervised learning tasks and has been studied for many years (Blum & Langley, 1997; Guyon & Elisseeff, 2003; Liu & Yu, 2005; Saeys et al., 2007).

Feature selection algorithms are believed to consist of two major components namely the Selection Criterion Function and the Search Strategy (Collins & Liu, 2005). The Selection Criterion Function is a quantitative measure which is used to compare one feature subset against another. The Search Strategy is a systematic method to enumerate candidate feature subsets and to decide when to stop (Collins & Liu, 2005). However, in this thesis, the author considered Stopping Criterion as a major component in addition to the two mentioned previously. While the Selection Criterion Function determines how to measure and quantitatively compare features and the Search Strategy provides the means to search for optimal feature subset through the feature subsets space, the Stopping Criterion determines if optimal or suboptimal feature subset is selected or attained. There are two categories of the Selection Criterion Functions which give rise to the three types of Feature Selection Models. Those that utilize a statistical and probability distribution of dataset attributes to measure relevance of feature subset and those that evaluate feature subset based on its performance with a learning machine model. The former referred to as Filter and the latter Wrapper. Hybrid model combines the strong points of both Filter and Wrapper.

1.2 Motivation

As the world is currently witnessing a sudden burst of data due to advances in technologies, the challenges of big data analytics become self-evident. As such the author of this thesis believes that effective feature selection methods will continue to play a role in uncovering most silent features in the data, that will make it easy for applying Machine Learning Algorithms to big data. Previous studies in (Kira & Rendell, 1992; Richeldi & Lanzi, 1996; Blum & Langley, 1997; Guyon & Elisseeff, 2003; Pudil & Somol, 2005; Chen & Lin, 2006; Rossi et al., 2006; Gadat & Younes, 2007; Doquire & Verleysen, 2011; Unler et al., 2011) highlighted that feature selection reduces the dimension of a dataset by selecting variables that are relevant to the predicting attribute; it helps to improve the predicting capability of machine learning models by building a compact model that often avoid over-fitting and generalized better, it can improve the accuracy of prediction because of reduction in estimation errors, and empirical results indicate that building good predictor model requires a decrease in feature subset. More so, feature subset selection reduces the burden of data collection as well as reducing computational complexity. The last two advantages of feature selection are of economic importance most especially in some domains that are capital-intensive and time-consuming to organize facilities and collect, or derived domain variables enough for a traditional empirical method of analysis. Feature selection will ensure the use of concise dataset to develop predictive model for a specific domain.

The two main types of feature selection criterion functions introduced in Section 1.1, each has its strong and weak points. Filter models are fast and select feature subset independent of any machine learning algorithm, as such it generalizes the use of selected feature subset by many machine learning algorithms. However, wrapper models are more efficient in choosing feature subset that gives better prediction model with a specific machine learning algorithm. Because