Faculty of Information and Communication Technology

IMPROVED RANDOM FOREST FOR FEATURE SELECTION IN WRITER IDENTIFICATION

Nooraziera Akmal Binti Sukor

Master of Science in Information and Communication Technology

2015

IMPROVED RANDOM FOREST TREE FOR FEATURE SELECTION IN WRITER IDENTIFICATION

NOORAZIERA AKMAL BINTI SUKOR

A thesis submitted
in fulfillment of the requirements for the degree of
Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2015

DECLARATION

I declare that this thesis entitled “Improved Random Forest Tree for Feature Selection in Writer Identification” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :……....

Name : Nooraziera Akmal Binti Sukor

Date :……....

APPROVAL

I hereby declare that I have read this thesis and my opinion this thesis is sufficient in term of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature : ……

Supervisor Name : PM Dr. Azah Kamilah Binti Draman @ Muda

Date : ……

DEDICATION

In the name of Allah swt, Beneficant, the Merciful
All praise is due to Allah, the Lord of the Worlds

I dedicated my dissertation work and give special thanks to:

MY PARENTS
(Haji Sukor Bin Awang&HajahZainonBintiHitam)

MY SUPERVISOR
(Prof. MadyaDr. AzahKamilahBintiDraman @ Muda)

MY FRIEND
(RimashadiraBintiRamlee)

MY SIBLINGS
(NoradilahBintiSukor, Noor HidayuBintiSukor, NadiyaSyakilaBintiSukor,
NurulNazihahBintiSukor&AnuarHafiez Bin Sukor)

for all encouragements and supports throughout the completion of this work.

ABSTRACT

Writer Identification (WI) is a process to determine the writer of a given handwriting sample. A handwriting sample consists of various types of features. These features are unique due to the writer's characteristics and individuality, which challenges the identification process. Some features do not provide useful information and may cause to decrease the performance of a classifier. Thus, feature selection process is implemented in WI process. Feature selection is a process to identify and select the most significant features from presented features in handwriting documents and to eliminate the irrelevant features. Due to the WI framework, discretization process is applied before the feature selection process. Discretization process was proven to increase the classification performances and improved the identification performance in WI. An algorithm and framework of Improved Random Forest (IRF) tree was applied for feature selection process. RF tree is a collection of tree predictors used to ensemble decision tree models with a randomized selection of features at each split. It involved Classification and Regression Tree (CART) during the development of tree. Important features are measured by using Variable Importance (VI). While Mean Absolute Error (MAE) values use to identify the variance between writers, VI value was used for splitting process in tree and MAE value is to ensure the intra-class (same writer) invariance is lower than inter-class (different writer) invariance because lower intra-class invariance indicates accuracy to the real author. Number of selected features and the classification accuracy is used to indicate the performances of feature selection method. Experimental results have shown that the performances of IRF tree in discretized dataset produced third feature (f3) as the most important feature with average classification accuracy 99.19%. For un- discretized dataset, first feature (f1) and third feature (f3) are the most important features with average classification accuracy 40.79%.

ABSTRAK

Pengenalpastian Penulis(PP) adalah satu proses bagi menentukan penulis berdasarkan sampel tulisan tangan yang diberikan. Sampel tulisan tangan terdiri daripada pelbagai jenis feature. Feature adalah unik berdasarkan ciri-ciri keperabadian penulis sehingga menimbulkan cabaran dalam proses pengenalpastian. Sebahagian dari feature adalah tidak relevan, tidak member maklumat yang berguna dan menyebabkan penurunan kadar prestasi pengelasan. Jadi, proses feature selection telah digunakan dalam proses PP. Feature selection adalah proses bagi mengenalpasti dan memilih feature yang paling penting daripada feature yang wujud dalam dokumen tulisan tangan dan membuang feature yang tidak relevan. Berdasarkan kerangka PP, proses discretization telah dicadangkan sebelum proses feature selection. Proses discretization telah terbukti meningkatkan persembahan klasifikasi dan memperbaiki persembahan pengecaman dalam PP. Algoritma dan kerangka Improved Random Forest (IRF) digunakan dalam proses feature selection. RF tree ialah koleksi pokok peramal menggunakan himpunan model decision tree, dengan memilih feature secara rawak pada setiap cabang. RF adalah salah satu kaedah embedded di mana ia melibatkan Classification and Regression Tree (CART) semasa pembangunan struktur pokok. Kepentingan feature diukur dengan menggunakan nilai Variable Importance (VI). Manakala nilai Min Absolute Error (MAE) digunakan untuk mengenalpasti variasi di antara penulis.Nilai VI digunakan untuk proses pemecahan pokok manakala MAE untuk memastikan variasi dari kumpulan penulis yang sama adalah lebih rendah dari variasi kumpulan penulis yang berbeza kerana nilai variasi yang rendah menunjukkan hamper kepada penulis sebenar. Bilangan feature yang terpilih dan ketepatan klasifikasi digunakan untuk mengukur persembahan kaedah feature selection. Keputusan eksperimen menunjukkan persembahan IRF tree dalam set data discretized telah menghasilkan feature ketiga (f3) sebagai feature yang paling penting dengan purata ketepatan klasifikasi 99.19%. Bagi dataset un- discretized, feature pertama (f1) dan feature ketiga (f3) adalah feature yang penting dengan purata ketepatan klasifikasi 40.79%.

ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude to my supervisor PM Dr. AzahKamilahBintiDraman@Muda for the continuous support for my Master study and research, for the patience, motivation, encouragement and immense knowledge. I would also express my gratitude to my co-supervisor, PuanNorazilahBintiDraman@Muda for the kind advice and valuable input.

My special thanks to my parentSukor Bin Awang and ZainonBintiHitam for giving birth to me at the first place and supporting me spiritually throughout my life. Also thankful to my siblings (NoradilahBintiSukor, Noor HidayuBintiSukor, NadiyaSyakilaBintiSukor, NurulNazihahBintiSukor and AnuarHafiez Bin Sukor) for all supports, encouragement and inspiration,

Last but not least, my sincere thanks also goes to my fellow friend, RimashadiraBintiRamlee for the stimulating discussions, friendship, for the sleepless nights we were working together before deadlines and for all the fun we had in the last seven years. I am grateful to everyone who has inspired me during the completion of this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HI	-	Handwriting Identification
HR	-	Handwriting Recognition
WI	-	Writer Identification
WV	-	Writer Verification
RF	-	Random Forest
VI	-	Variable Importance
MSE	-	Mean Squared Error
OOB	-	Out-Of-Bag
CART	-	Classification and Regression Tree
UMI	-	United Moment Invariant
SFS	-	Sequential Forward Selection
SFFS	-	Sequential Forward Floating Selection
CI-SFFS	-	Computationally Inexpensive Sequential Forward Floating Selection
MIC	-	Modiffied Immune Classifier
CFS	-	Correlation-based Feature Selection
LVF	-	Las Vegas Filter
FCBF	-	Fast Correlation-based Filter
SMFS	-	Significance Measurement Feature Selection
NSA	-	Negative Selection Algorithm
NBayes	-	Naïve Bayes
SVM	-	Support Vector Machine

LIST OF SYMBOLS

$\hat{y}$	-	Training set
$T(x)$	-	Tree
M	-	Bagging
$\hat{c}_m$	-	Prediction Error
$\hat{R}_m$	-	Impurity Measurement
N_m	-	Number of Sample
$\hat{f}$	-	Prediction
B	-	Number of Tree
x'	-	Sample Data
x_p	-	Predictor
$\hat{e}_k$	-	Prediction Error
$\hat{e}$	-	Out-Of-Bag Error
$\hat{y}_i^{OOB}$	-	average prediction for i th observation
MSE_{OOB}	-	squared error for OOB
$\hat{\sigma}_y^2$	-	standard deviation taken over tree computed with n as advisor
n	-	number of split branches in tree

LIST OF APPENDICES

LIST OF PUBLICATIONS

SUKOR N.A., MUDA, A. K., MUDA, N. A.&CHOO, Y.-H 2014 A Comparative Study of Tree-based Structure Methods for Handwriting Identification.*First International Conference on Advanced Data and Information Enginerring (DaEng-2013) Lecture Notes in Electrical Engineering* Volume 285, 2014, pp 269-276

SUKOR N.A., MUDA, A. K., MUDA, N. A.&CHOO, Y.-H 2014 Tree-base Structure for Feature Selection in Writer Identification. *Pattern Analysis, Intelligent Security and the Internet of Things (2015) Springer International Publishing*pp 201-213

CHAPTER 1

INTRODUCTION

1.0 Introduction

Writer Identification (WI) problems was introduced long time ago and various studies were conducted using pattern recognition technique (Srihari*et al.,* 2001; Schlapbach and Bunke, 2004; Zhang and Srihari, 2003; Zhu *et al*, 2004) where the result is to identify the author of a handwritten document. The uniqueness and the individuality of a handwriting style can be used to identify the significant features in identifying the original owner of the handwriting. There are some applications that required WI performances such as identifying the writer on legal papers by signature, handling threat letters or determination of an old or historical manuscript. WI studies contribute a great importance towards the criminal justice system and have been widely explored in forensic handwriting analysis (Somaya*et al.,* 2008; Srihari*et al.,* 2006)

One handwritten document presents various types of features. These features refer to the writer's characteristics or individuality (Muda*et al.*, 2007), which analyzes the shape of character and words combined (Huber and Headrick, 1999). The irrelevant and redundant information will only decrease the performance of a classifier. Removing these irrelevant and redundant features can improve the classification accuracy (Muda, 2009). This study is focused on the feature selection process which is a process to select significant features among the presented features in handwritten documents. These selected features have a major impact in identify the writer during WI process

1.1 Background of the Study

Handwriting Analysis is divided into Handwriting Identification (HI) and Handwriting Recognition (HR). HI is used to identify the writer of the given handwritten document while HR deals with the content and meaning of handwritten text. There are two models in HI which are writer identification (WI) and writer verification (WV), as shown in Figure 1.1. WI is the process to determine who is the writer among the candidates for the given handwriting samples (focused in this study) while WV deals with a given samples of handwriting and then determine whether the sample belongs to the same writer or different writer among the candidates

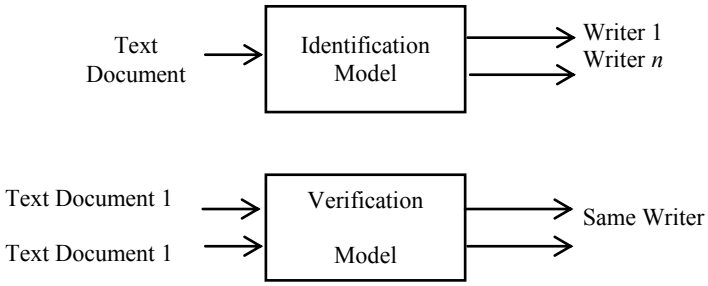

Figure 1.1: Handwriting Identification Model

Handwriting has an individualistic nature where each individual has a natural variation where no two people have the exact handwriting style (Srihari*et al.,* 2006). In the forensic field, handwriting is one of evidence that is most involved in cases such as forgery, murder and etc. This is because HI is a personal biometric attribute and is considered unique to every person (Srihari*et al.*, 2001; Srihari, *et al.*, 2002; Zhang and Srihari, 2003). The shapes and the style of writing can be used as biometric features for authenticating and identifying a person (Srihari*et al.*, 2006).

Each of the features is different due to the various handwriting styles and some do not provide useful information in WI. This will in turn lead to interrupting the performance of the classifier, thus creating a challenge in the identification process. To overcome this problem, feature selection process will be applied to select the most significant of features thus increasing the classification accuracy (Muda *et al.,* 2011). This is due to the feature selection process where only significant features will be used in the classification process and thus can minimize the complexity processing while improving the classification accuracy. This leads to the main issue in Writer Identification which is how to acquire the most significant of features reflected in the author's handwriting (Srihari *et al.,* 2002; Shen *et al.,* 2002).

There have been many studies done on feature selection process in handwriting's field such as: Feature Selection using Genetic Algorithm for Handwritten Character (G. Kim and S. Kim, 2000), Feature Selection for ensemble applied to Handwritten Recognition (L.S. Oliveira, 2006), A Feature Selection Algorithm for Handwritten Character Recognition (L. Cordella, 2008), Feature Selection for Recognizing Handwritten Arabic Letters (A. Gheith *et al*., 2010) and a GA-based Feature Selection Approach with an application to Handwritten Character Recognition (C. De Stefano *et al.*, 2013). Also the research of Comparative Study of Feature Selection Method for Authorship Invariances in Writer Identification (Pratama S. F *et al.*, 2012) and Comparative Study of Feature Selection Method for Writer Identification (Pratama S. F *et al.*, 2013)

There are three popular methods of feature selection which are filter method, wrapper method, and embedded method (Saeys *et al.,* 2007). Based on the literature review, most of the researches involve filter method and wrapper method for feature selection in handwriting domain. Very little exploration has been done for embedded

method in the handwriting domain. This study will explore how embedded method works on feature selection process in WI and focuses on tree-based structure.

1.2 Statement of the Problem

Handwriting is individualistic where every person has his/her own handwriting styles. All features from following criteria affect the individuality of handwriting which makes ones' handwriting differ from another:

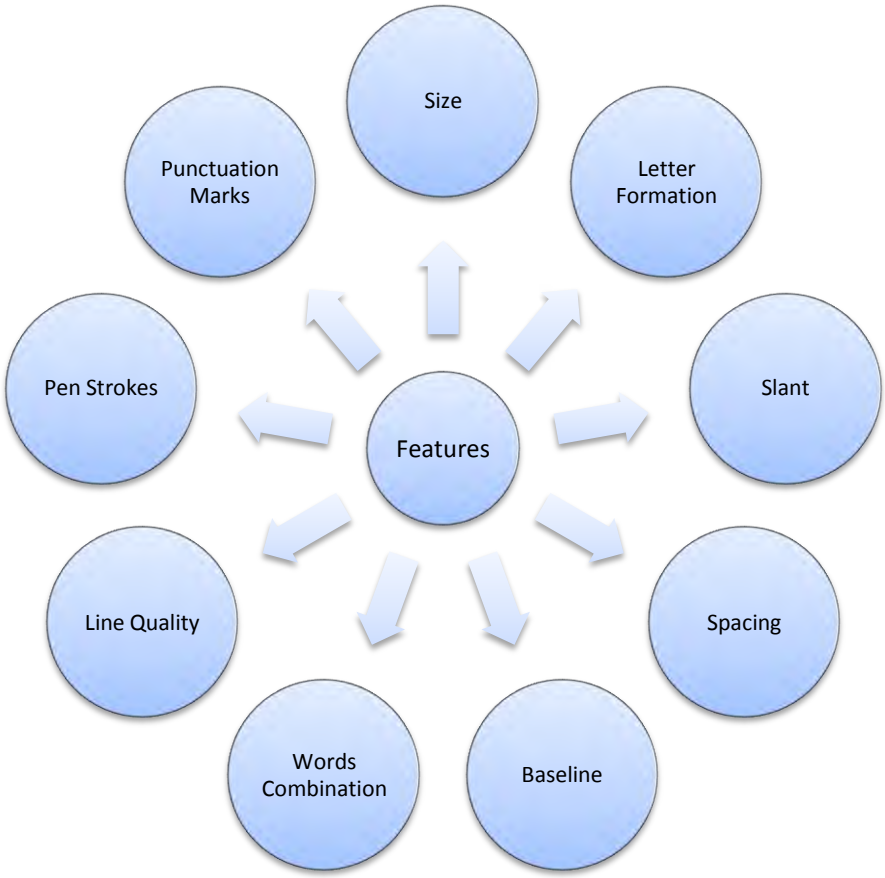

Figure 1.2: Handwriting Features

Some The primary problem statement for this research is:

"How to obtain the effective tree structure in selecting the significant and unique features of handwriting for Writer Identification?"

In order to complement the primary problem statement, there are three secondary problem statements to be considered:

1) How to identify the significant features in Writer Identification?

2) How to ensure the effective framework for feature selection in Writer Identification process?

3) How to verify the proposed feature selection method is capable in selecting significant features?

1.3 Objectives of the Study

Some features are relevant to classification process and some are not. Irrelevant features will interrupt the classification process and increase the complexity of WI process (Pratama *et al.,* 2007). Hence, feature selection process is important to eliminate these irrelevant features. During feature selection process, only relevant features will be obtained. Relevant features are unique to individual and give a major impact to the one's handwriting. The performances of feature selection process will be measured against the number of selected features and classification accuracy. Time complexity is excluded as it is not an issue in WI domain (Pratama*et al.,* 2007).

The embedded method of feature selection is still limited use in WI domain. Tree-based structure is an example of embedded method and Random Forest (RF) tree will be further explored in this study. In previous study, RF tree is used for classification process in WI. This study will be applying RF tree method for feature selection process in WI with

the purpose of producing minimal subset of features while giving high classification accuracy. It can be achieved with the following objectives:

1. To propose an algorithm of tree structure-based for feature selection in WI.
2. To propose a framework of tree structure-based technique for feature selection in WI.

1.4 Assumptions and Hypothesis

During this study, an experiment regarding WI was carried out. This experiment involved a dataset from IAM handwriting database as it is a benchmarked dataset used by other researchers in the handwriting field.

Filter method, wrapper method and embedded method are the most popular feature selection methods that are usually implemented in WI domain. However, in this study tree-based structure of embedded method will be further explored and Random Forest (RF) tree is selected as feature selection method in WI process. RF tree works by scoring the importance features and find the absolute error of features using method of Variable Importance (VI) and Mean Absolute Error (MAE). Relevant features means high value of VI. In WI, the variance for intra-class (same writer) must be lower than inter-class (different writer) (He *et al.*, 2008; Srihari *et al.*, 2001; Zois and Anastassopoulos, 2000) which is indicates accuracy to the real author

The hypotheses of the study is that RF tree is capable to select only relevant features (depends on VI) during feature selection process and produces high percentage of classification accuracy in WI.