FEATURE SELECTION USING FAST CORRELATION-BASED FILTER (FCBF) FOR LEAF CLASSIFICATION

LUA XIN LIN

This report is submitted in partial fulfilment of the requirements for the Bachelor of Computer Science (Artificial Intelligence)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2015

BORANG PENGESAHAN STATUS TESIS*

JUDUL: **FEATURE SELECTION USING FAST CORRELATION-BASED FILTER (FCBF) FOR LEAF CLASSIFICATION**

SESI PENGAJIAN: **2014/2015**

Saya **LUA XIN LIN** mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat
3. salinan untuk tujuan pengajian sahaja.
4. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat
5. salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
6. ** Sila tandakan (/)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
_/	TIDAK TERHAD	

(TANDATANGAN PENULIS)

ALAMAT TETAP: 49 Jln BU4/9
Bandar Utama 47800 Petaling Jaya,
Selangor Darul Ehsan, Malaysia
Tarikh:

(TANDATANGAN PENYELIA)

(NAMA PENYELIA)

Tarikh:

CATATAN:
* Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

DECLARATION

I hereby declare that this project report entitled

FEATURE SELECTION USING FAST CORRELATION-BASED FILTER (FCBF) FOR LEAF CLASSIFICATION

is written by me and is my own effort and that no part has been plagiarized without citation.

STUDENT: _____ DATE: _____

(LUA XIN LIN)

SUPERVISOR: _____ DATE: _____

(DR AZAH KAMILAH BINTI DRAMAN @ MUDA)

DEDICATION

This work is dedicated to my family and friends. To my parents, for your unwavering support—mom and dad, for always nagging at me to get on with my work. My brother, Alex, for your sarcasm and wit—really, what are siblings for if not some friendly squabbling? To my dear friend Zoe, for not begrudging my late night calls, patiently listening to all my woes and the many words of encouragement—you're my rock. To my supervisor, Dr Azah Kamilah bt. Draman@Muda, for making all this possible.

ACKNOWLEDGEMENT

I'd like to thank my supervisor, Dr Azah Kamilah bt Draman@Muda, who has been more than generous with her expertise and time. Your guidance is priceless.

I'd also like to thank Dr Choo Yun Huoy, for willingly imparting advice and information. I immensely enjoyed your Machine Learning lectures; those lessons have aided me well in the process of this work.

ABSTRACT

Imagine taking a forest trail hike, where you see many interesting plants. How would you differentiate one plant from another based on its leaves? Do you base your decision on its color? Or its texture? Or size? The thing is, one single leaf could give you a lot of information, from color to texture to size, so on and so forth. But the problem is: which features to focus on so as to make a good decision?

The identification of a feature subset that best represents a class so as to build a strong predictive model is still an issue that researchers are still working on solving. This paper focuses on feature selection, where the Fast Correlation-Based Filter is compared to the Correlation-based Feature Selector.

Four datasets were retrieved from the UCI Machine Learning Repository. The SVM classifier was used to build a predictive model based on the feature subset. The experiments were done using the weka machine learning tool, with the evaluation criterion being predictive accuracy, kappa measure, and time taken to build a model.

ABSTRAK

Bayangkan, anda sedang trek di hutan, di mana anda melihat banyak tumbuhan yang menarik. Bagaimana anda membezakan satu tumbuhan dari yang lain berdasarkan daun? Adakah anda mendasarkan keputusan anda pada warnanya? Atau tekstur? Atau saiz? Masalahnya, satu daun tunggal boleh memberikan banyak maklumat, dari segi warna, tekstur, saiz dan sebagainya. Tetapi masalahnya ialah: ciri-ciri manakah hang haru diberi tumpuan supaya keputusan yang baik dapat dilakukan?

Pengenalpastian subset 'feature', ataupun ciri yang terbaik mewakili kelas bagi membina model ramalan yang kuat masih merupakan isu yang penyelidik masih berusaha untuk menyelesaikan. Tesis ini memberi tumpuan kepada pemilihan 'feature', di mana Fast Correlation Based Filter dibandingkan dengan ciri Correlation-Based Filter.

Empat set data telah diperolehi dari UCI Machine Learning Repository. Classifier SVM telah digunakan untuk membina model ramalan berdasarkan subset 'feature' tersebut. Eksperimen telah dilakukan dengan menggunakan alat pembelajaran mesin weka, dengan kriteria penilaian yang berikut: ketepatan ramalan, statistik kappa, dan masa yang diambil untuk membina model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

FS	--	Feature Selection
PSO	--	Particle Swarm Optimization
BFS	--	Best First Search
SVM	--	Support Vector Machine
GCLM	--	Grey-Level Co-concurrence Matrix
SMO	--	Sequential Minimal Optimization
KKT	--	Karush-Kuhn Tucker
FCBF	--	Fast Correlation-Based Filter
CFS	--	Correlation-Based Feature Selection
PNN	--	Probabilistic Neural Networks
kNN	--	k-Nearest Neighbor
GA	--	Genetic Algorithm
PCA	--	Primciple Component Analysys

CHAPTER 1

INTRODUCTION

1.1 Introduction

One of the upcoming fields in need of research and development is the automatic recognition of plant species—plant classification. This is due to the vital role plants play in the natural circle of life: the conversion of carbon dioxide to oxygen. Despite this being general knowledge, deforestation continues at large with little consideration to its ramifications, which will result in the imminent extinction of a large number of plant species. According to earthsendangered.com, there are approximately 9000 endangered plant species worldwide, as of March 2015. This staggering figure necessitates establishing a plant species database dedicated to plant protection and data preservation.

Prior to the advances in computing technology, there exist a number of well-established plant classification methods. One of the earliest and most well-known methods is the Linnaeus taxonomy, whereby plants were subdivided into 24 orders of classes, which were used to identify but not represent natural groups of plants. At present, plant taxonomy still largely adopts the traditional classification methods.

Other modern methods of plant classification—molecular biology and morphological anatomy, to name a few—are deeply rooted in biology and chemistry.

The introduction of multidisciplinary studies such as image processing and machine learning in the computer science sector in the past decade has led to research in this field, producing non-manual plant classification systems. Compared to the abovementioned techniques, this method may prove superior, as leaf sampling and photographing said samples are less costly. The images of leaves can also be easily transferred to a computer so that the features can be extracted via image processing techniques.

The challenge, however, lies in the extraction and identification of discriminant features distinguishing plant features. While an image of a leaf will provide an array of data to be used with classification algorithms, not all the extracted features are relevant –in fact, too many irrelevant features inhibit the accuracy of results. Therefore, it is vital to identify relevant features in order to produce a classification model that can accurately classify a plant based on input vectors.

1.2 Problem statement

As mentioned above, a lot of data, especially redundant data will cause the predicted output to deviate from the actual result. As such, feature selection is a must in the task of automatic plant classification in order to effectively eliminate irrelevant attributes, which minimizes the errors in classification.

In the past, multiple feature selection techniques have been known to work well, such as the PSO and ACO. However, the recent increase in the dimensionality of data poses a severe challenge to many existing feature selection techniques with respect to effectiveness and efficiency.

In the domain of agricultural robotics, automatically distinguishing between plant species is a challenging task, especially because some species are physically very similar to each other. The number of attributes one can use to classify a plants' species is thus too vast, whereby the use of an irrelevant/redundant attribute in the classification process may result in inaccurate results.

This study is done with the aim of identifying the most appropriate method of feature selection that applies to plant classification using SVM.

1.3 Objectives

- To study the efficiency of the Fast Correlated Based Filter(FCBF) feature selector algorithm in respect to predictive accuracy of the SVM in plant classification
- To propose the use of the Fast Correlated Based Filter as a feature selector for plant classification using SVM, if viable
- To compare the efficiency of the FCBF with other feature selecting algorithms.
- To evaluate the proposed techniques in feature selection for plant leaf classification

1.4 Scope

- Focus of project : plant leaf classification
- Experiments run using the Weka machine learning tool

  - 3 feature selection algorithm (FCBF, CFS with PSO, and CFS with Best-First Search)
  - The SMO classifier with Poly-kernel
- 2 sets of data:
  - Dataset 1 consists of 3 datasets, each with 1600 instances and 64 feature vectors: margin, texture and edge
  - Dataset 2 consists of 340 instances and 16 attributes

1.5 Project Significance

- Pinpoint the suitability of proposed feature selection/preprocessing for the chosen datasets
- Qualitative analysis: guideline/reference for future work by other researchers

1.6 Expected results

- Analysis of classification accuracy of SVM on features selected using abovementioned techniques
- Analysis of performance of feature selection technique

1.7 Conclusion

The FCBF feature selector will be compared to two other feature selectors in the task of plant leaf classification in this project. The next chapter will cover literature reviews on this topic.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Much research has been done on the application of Machine Learning in the agricultural field of plant leaf identification. While this chapter covers a variety of such application, this review focuses on the machine learning aspect of the process of plant leaf classification. The processes involved in plant leaf identification are: a) image acquisition, the acquisition of leaf images; b) Feature extraction: the extraction of data from leaf images; c) Feature selection: the selection of relevant, non-redundant features; d) Classification: the prediction of a plant species based on an array of attributes selected by the feature selector. Although this literature extensively covers the process of plant leaf identification, this paper will primarily focus on feature selection.

2.2 Plant Leaf Identification

The importance of an automated classification system for plants has been detailed in the previous chapter (chapter one).

For the past few years, researchers have worked on various techniques of automated plant classification. Generally, the process of classifying plants is as shown in Figure 2.1.

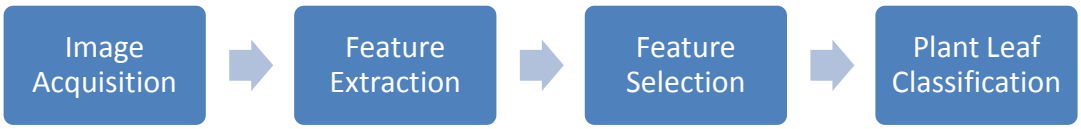

Figure 2.1: Process of Plant Leaf Classification

In *A leaf recognition algorithm for plant classification using probabilistic neural network* Wu et.al. (2007) proposed the probabilistic neural network (PNN) as a plant classification algorithm. In this work, 12 features were extracted and orthogonalized via Principle Component Analysis (PCA) to produce 5 principle components, which serve as inputs for the PNN. This method resulted in an accuracy of 90.312% when used on the Flavia dataset (which was provided by We et.al and available to the public). However, probabilistic neural networks are usually applied on smaller benchmarking datasets, as too large an array of inputs with which to train the network may result in overfitting. In *Neural networks for classification: a survey*, Zhang (2000) details the overfitting effect on test sets when using Probabilistic Neural Networks.

In *Svm-bdt pnn and fourier moment technique for classification of leaf shape*, Singh et.al. (2010) presented three methods of leaf classification: the SVM-BDT (Support Vector Machine utilizing Binary Decision Tree), PNN with PCA and Fourier moment, all of which utilized the same dataset as Wu et.al. (2007). In this research, the authors found that the SVM was superior to the PNN and Fourier moment due to its high generalization ability without the need of *a priori* knowledge.