



Tribhuvan University

Institute of Science and Technology

HYBRID FEATURE SELECTION AND FEATURE EXTRACTION
BASED ENSEMBLE METHOD IN CLASSIFICATION

A Dissertation

Submitted to:

Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur

In partial fulfillment of the requirements
for the Master's Degree in Computer Science & Information Technology

Submitted by:

Rajesh Pandey
Sept, 2015

Supervisor

Prof. Dr. Shashidhar Ram Joshi



Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information
Technology

Student's Declaration

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

.....

Rajesh Pandey

Date: 22th Sept, 2015



Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information
Technology

Supervisor's Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Rajesh Pandey** entitled “**Hybrid feature selection and feature extraction based ensemble method in classification**” be accepted as in fulfilling partial requirement for the completion of Masters Degree of Science in Computer Science & Information Technology.

Prof. Dr. Shashidhar Ram Joshi

Department of Electronics & Computer Engineering,

Institute of Engineering,

Pulchowk, Nepal

Date: 23th Sept, 2015



Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information
Technology

LETTER OF APPROVAL

We certify that we have read this dissertation work and in our opinion it is appreciable for the scope and quality as a dissertation in the partial fulfillment of the requirements of Masters Degree of Science in Computer Science & Information Technology.

Evaluation Committee

Asst. Prof. Nawaraj Paudel
Head of Department
Central Department of Computer Science
& Information Technology
Tribhuvan University
Kirtipur

Prof. Dr. Shashidhar Ram Joshi
Department of Electronics & Computer
Engineering,
Institute of Engineering,
Pulchowk, Nepal

(External Examiner)

(Internal Examiner)

Date: 23 Dec, 2015

Acknowledgement

I would never have been able to finish my dissertation without the guidance, support and encouragement of numerous people including my supervisor, my friends, colleagues and support from my family. At the end of my thesis I would like to thank all those people who made this thesis possible and an unforgettable experience for me.

First, I would like to express my gratitude to my supervisor **Professor Dr. Shashidhar Ram Joshi**, Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk for his continuous support without which the thesis wouldn't have been possible to complete. His suggestions, guidance, thorough knowledge and expertise helped me immensely in understanding and developing this thesis. I thank him immensely for his patience and generous time spent to guide me through the entire process.

Most importantly I would like to thank to respected Head of Department of Central Department of Computer Science and Information Technology, Asst. Prof. Nawaraj Paudel for his kind support, help and constructive suggestions. I am very much grateful and thankful to all the respected teachers Prof. Dr. Subarna Sakya, Dr. Arun Kumar Timilsina, Mr. Min Bahadur Khati, Mr. Dheeraj Kedar Pandey, Mr. Sarbin Sayami, Mrs. Lalita Sthapit, Mr. Arjun Singh Saud, Mr. Bikash Balami and Tej Bahadur Shahi for providing me such a broad knowledge and inspirations.

Special thanks to my family for their endless motivation, constant mental support and love which have been influential in whatever I have achieved so far. All my class fellows are worthy of my gratefulness for their direct or indirect support in completion of my dissertation. Finally, I would like to thank to respected teacher Mr. Jagdish Bhatt for his suggestions during my work.

I have done my best to complete this research work. Suggestions from the readers are always welcomed, which will improve this work.

Abstract

Ensemble methods have been called the most influential development in Data Mining and Machine Learning in the past decade. The idea of ensemble learning is to employ multiple learners and combine their predictions.

In this thesis, a novel method is proposed to build an ensemble of classifiers based on feature selection: Random selection, Relief and feature extraction: Principal component analysis method. The feature selection process chooses optimal subset of features according to objective function whereas feature extraction process maps the high dimensional dataset into lower dimensional dataset using the linear combination of original features. These feature selection and extraction method helps to produce diverse as well as accurate set of ensemble classifiers. A comparison of proposed method is made with the Bagging, AdaBoost, feature selection based NN, feature extraction based NN and also with plain NN using 22 benchmark dataset. The result obtained by the proposed method outperformed other algorithms with the following distribution: NN (14 cases), Random-NN (13 cases), Relief-NN (15 cases), PCA-NN (19 cases), AdaBoost (14 cases), Bagging (15 cases).

Keywords: Ensemble methods, feature selection, feature extraction, Relief, Principal component analysis, AdaBoost, Bagging, NN, Random-NN, Relief-NN, PCA-NN

Table of Contents

<u>Contents</u>	<u>Page No.</u>
Abstract	i
List of Figure	v
List of Abbreviations	vi
Chapter 1: Introduction	1
1.1 Introduction	1
1.2 Thesis Organisation.....	1
Chapter 2: Background study & Problem Formulation	3
2.1 Ensemble method	3
2.2 Dimension Reduction	4
2.2.1 Feature selection	5
2.2.1.1 Filters.....	5
2.2.1.2 Wrappers	6
2.2.1.3 Embedded	6
2.2.2 Feature Extraction	6
2.3 Problem Formulation	7
2.3.1 Problem Statement.....	7
2.3.2 Objectives	7
2.4 Motivation.....	7
Chapter 3: Literature Review & Methodology	9
3.1 Literature Review	9
3.1.1 Feature selection and Extraction	9
3.1.2 Ensemble Method	10
3.2 Methodology	14
3.2.1 Research Methodology	14
3.2.2 Hybrid Feature selection and Extraction based Ensemble method	14
3.2.2.1 K-NN Classifier	15
3.2.2.2 Random feature selection	15
3.2.2.3 Relief Feature selection method	15
3.2.2.4 Principal Component Analysis	17
3.2.2.4.1 Best line Approximation	17
3.2.2.5 Majority Vote	19

3.2.2.5.1 Accuracy of the Majority Vote	20
3.2.2.6 Construction of HFEE Method	20
3.2.2.6.1 Training Phase.....	21
3.2.2.6.2 Testing Phase.....	21
Chapter 4: Implementation	23
4.1 Tools used	23
4.1.1 Programming language	23
4.1.2 Eclipse IDE	23
4.1.3 Weka Workbench	24
4.2 PCA module	24
4.3 Relief module	25
4.4 Majority Voting Algorithm	26
Chapter 5: Data Collection & Analysis	28
5.1 Data Collection	28
5.1.1 Training & Testing data	29
5.2 Experiment & Result	30
5.2.1 Experimental setup	30
5.2.2 Evaluation metrics	30
5.2.2.1 Precision	31
5.2.2.2 Recall	33
5.2.2.3 F-measure	34
5.2.2.4 Averages of Evaluation metrics	36
5.2.3 Final Result	37
Chapter 6: Conclusion & Future Work	39
6.1 Conclusion	39
6.2 Future Work	39
References	40
Bibliography	41
Appendix	42

List of Figures

Figure 2.1: Ensemble Classifier.....	4
Figure 3.2.2.6.1: Ensemble training phase.....	21
Figure 3.2.2.6.2: Ensemble testing phase.....	22
Figure 5.1.1 (a): Sample data of Dermatology dataset used for training	29
Figure 5.1.1 (b): Sample data of Credit dataset used for training	29
Figure 5.1.1 (c): Sample data of Dermatology dataset used for testing.....	29
Figure 5.1.1 (d): Sample data of Credit dataset used for testing	30
Figure 5.2.2.1: Graph showing Average precision of Algorithm.....	32
Figure 5.2.2.2: Graph showing Average recall of Algorithms	34
Figure 5.2.2.3 Graph showing Average F-measure of Algorithms	36
Figure 5.2.2.4: Graph showing averages of Evaluation metrics	37

List of Tables

Table 5.1: List of Data	28
Table 5.2.2.1: Result taking Precision	32
Table 5.2.2.2: Result taking Recall	34
Table 5.2.2.3: Result taking F-measure	35
Table 5.2.2.4: Result taking Averages of Evaluation metrics	36

List of Abbreviations

Abbreviations

NN

PCA

EOF

PC

CFS

MDL

HFEE

WEKA

SDK

IDE

SWT

Full Form

Nearest Neighbor

Principal Component Analysis

Empirical orthogonal function

Principal component

Correlation feature selection

Minimum Description Length

Hybrid feature selection & Extraction Ensemble

Waikato Environment for Knowledge Analysis

Software development kit

Integrated development Environment

Standard Widget toolkit