EXTENDED NEAREST CENTROID NEIGHBOR METHOD WITH TRAINING SET REDUCTION FOR CLASSIFICATION

NORDIANA BINTI MUKAHAR

UNIVERSITI SAINS MALAYSIA

2020

EXTENDED NEAREST CENTROID NEIGHBOR METHOD WITH TRAINING SET REDUCTION FOR CLASSIFICATION

by

NORDIANA BINTI MUKAHAR

Thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

JUNE 2020

ACKNOWLEDGEMENT

All praise to ALLAH S.W.T the Almighty for giving me the opportunity, strength, patience and ability to complete my study at Universiti Sains Malaysia. I would like to express my sincere gratitude to my supervisor, Associate Professor Dr. Bakhtiar Affendi bin Rosdi for his continuous encouragement, guidance and support throughout this research. Thank you for your informative suggestion and brilliant ideas, constructive and valuable comments and for giving me the flexibility to work at my own pace.

This thesis is dedicated to all my beloved: especially my mother, Pn. Hajjah Salmah Binti Mustajab. Her continuous support, love and prayer make me strong to endure this voyage. My heartfelt thanks to my beloved husband, Mohd Nazri Bin Amir Hamezah and my wonderful daughter, Nusaibah for always being there with me to encourage and care and for understanding and giving me the strength to face this journey. I am greatly indebted and appreciate very much to my husband for his patience, sacrifice and support in my life.

My gratitude is also expressed to all my friends and postgraduate roommates for their continuous support, help and understanding throughout this period. I would like to extend my appreciation to my main sponsor, the Ministry of Higher Education, Malaysia and my employer, Universiti Teknologi MARA for supporting me in finishing this study. I would also want to thank the administrative and technical staff of the School of Electrical and Electronic Engineering USM and Institute of Postgraduate Studies (IPS) for their kind support and assistance.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENT ii

TABLE OF CONTENTS iii

LIST OF TABLES viii

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xvii

LIST OF SYMBOLS xix

ABSTRAK xxiii

ABSTRACT xxv

CHAPTER 1 - INTRODUCTION

1.1 Overview of Non-parametric Classifiers 1

1.2 Problem Statement and Motivation 3

1.2.1 Slow Classification Time of the *k* - Nearest Centroid Neighbor Classifier 4

1.2.2 One-sided Selection of Nearest Centroid Neighbors 6

1.3 Research Objectives 12

1.4 Scopes of Research 13

1.5 Thesis Outlines 14

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction 16

2.2 Non-parametric Classifiers 16

2.3 Nearest Neighbor Based Classifiers 17

2.3.1 *k* Nearest Neighbor Classifier 17

2.3.2 Weighting Scheme in the *k* - Nearest Neighbor 23

Classifier

2.3.3 Two-sided Selection of Nearest Neighbors in the *k* - Nearest Neighbor Classifier 25

2.4 Nearest Centroid Neighbor Based Classifiers 40

2.4.1 *k* - Nearest Centroid Neighbor Classifier 41

2.4.2 Variants of the *k* - Nearest Centroid Neighbor Classifiers 47

2.5 Training Set Reduction Techniques of the *k* - Nearest Centroid Neighbor Classifier 52

2.5.1 Decremental Techniques of the *k* - Nearest Centroid Neighbor Classifier 53

2.5.2 Limiting the Training Set of the *k* - Nearest Centroid Neighbor Classifier. 58

2.6 Summary 68

CHAPTER 3 – REDUCED SET *k* - NEAREST CENTROID NEIGHBOR CLASSIFIERS

3.1 Introduction 70

3.2 Training Set Reduction Techniques 71

3.2.1 Reduced Set *k* - Nearest Centroid Neighbor.v1 72

3.2.2 Reduced Set *k* - Nearest Centroid Neighbor.v2 82

3.2.3 Reduced Set *k* - Nearest Centroid Neighbor.v3 90

3.2.4 Reduced Set *k* - Nearest Centroid Neighbor.v4 98

3.3 The Differences of the Variants of the Reduced Set *k* - Nearest Centroid Neighbor, Wilson's Edited kNCN, Iterative kNCN and Limited-kNCN Techniques 105

3.4 Performance Evaluation 125

3.5 Summary 127

CHAPTER 4 – EXTENDED NEAREST CENTROID NEIGHBOR CLASSIFIER WITH REDUCED TRAINING SET

4.1 Introduction 130

4.2 Extended Nearest Centroid Neighbor Classifier 131

4.2.1 Procedure of the ENCN Classifier 134

4.2.2 The Differences between the kNCN and ENCN Classifiers 140

4.3 Reduced Set Extended Nearest Centroid Neighbor Classifier 149

4.3.1 Procedure of the RSENCN Classifier 154

4.3.2 The Differences between the ENCN and RSENCN Classifiers 157

4.4 Performance Evaluation 162

4.5 Summary 163

CHAPTER 5 – EXPERIMENTAL PROCEDURES, RESULTS AND DISCUSSION

5.1 Introduction 166

5.2 Databases 168

5.2.1 Real-world Database: UCI Machine Learning and KEEL Databases 169

5.2.2 Artificial Data Set: I-4I Data Set 171

5.2.3 Finger Vein Universiti Sains Malaysia (FV-USM) Images Database 172

5.3 Results of the Training Set Reduction Techniques 173

5.3.1 Selection of the Parameter, α for the RSkNCN.v2 and RSkNCN.v4 Techniques. 173

5.3.1 (a) Experiment 1: Performance Evaluation of the RSkNCN.v2 and RSkNCN.v4 174

Techniques on the Real-world Data Sets

5.3.1 (b) Experiment 2: Performance Evaluation of the RSkNCN.v2 and RSkNCN.v4 Techniques on the FV-USM Image Database 177

5.3.2 Performance of the Training Set Reduction Techniques 180

5.3.2 (a) Performance of Classification on the Real-world Data Sets 180

5.3.2 (b) Performance of Classification on the FV-USM Image Database 191

5.4 Results of the Reduced Set Extended Nearest Centroid Neighbor Classifier 196

5.4.1 Selection of the Training Set Reduction Technique for the ENCN Classifier 196

5.4.1 (a) Experiment 1: Performance Evaluation of the ENCN Classifier with Four Variants of the Training Set Reduction Techniques on the Real-world Data Sets 197

5.4.1 (b) Experiment 2: Performance Evaluation of the ENCN Classifier with Four Variants of the Training Set Reduction Techniques on the FV-USM Image Database 206

5.4.2 Performance of the RSENCN Classifier 210

5.4.2 (a) Performance of Classification on the Real-world Data Sets 210

5.4.2 (b) Performance of Classification on the I-4I Data Set 221

5.4.2 (c) Performance of Classification on the FV-USM Image Database 224

5.5 Summary 228

CHAPTER 6 - CONCLUSIONS AND FUTURE WORKS

6.1 Conclusions 230

6.2 Thesis Contributions 233

6.3 Suggestions and Future Works 233

REFERENCES 235

APPENDIX A Performance of the training set reduction techniques in terms of classification accuracy (%), reduction ratio, classification time (s) and learning time (s) on the Real-world data sets and FV-USM image database

APPENDIX B Performance of the combination ENCN classifier with four training set reduction techniques in terms of classification accuracy (%), reduction ratio and classification time (s) on the Real-world data sets and FV-USM image database

APPENDIX C Classification accuracy of the RSENCN classifier via different values of *k* on the Real-world data sets

LIST OF PUBLICATIONS

LIST OF TABLES

Page

Table 2.1 Sets of k - nearest neighbors for each training sample of the classes c_1 and c_2 before the introduction of the test sample, x in an offline stage (k = 3). 36

Table 2.2 New sets of k - nearest neighbors for each training sample of the classes c_1 and c_2 after the introduction of the test sample, x in an online stage ($k = 3$). 38

Table 2.3 New sets of k - nearest neighbors for each training sample of the classes c_1 and c_2 after the introduction of the test sample, x in an online stage ($k = 3$). 39

Table 2.4 Summary of the advantages and disadvantages of the kNN-based and kNCN-based classifiers. 50

Table 2.5 Sets of k - nearest neighbors for training samples of the classes c_1 and c_2 in the training set, T. ($k = 3$). 55

Table 2.6 Sets of k - nearest neighbors for training samples of the classes c_1 and c_2 in the subset, S (obtained from the first iteration) on the second iteration. 57

Table 2.7 Sets of k - nearest neighbors for training samples of the classes c_1 and c_2 in the subset, S (obtained from the second iteration) on the third iteration. 58

Table 2.8 Steps to determine the maximum rank of k - nearest centroid neighbors of the training sample, y_2 (m_{rank,y_2}) when $k = 3$. 63

Table 2.9 Summary of the advantages and disadvantages of the kNCN-based training set reduction techniques. 66

Table 3.1 Classification accuracy (%) obtained by the RSkNCN.v2 on the Wdbc data set when varying the parameter, α for k = 3. 111

Table 3.2 Classification accuracy (%) obtained by the RSkNCN.v4 on the Wdbc data set when varying parameter, α for k = 3. 113

Table 3.3 Comparisons in terms of the size of the subset obtained by various training set reduction techniques. 116

Table 3.4 Comparisons in terms of the classification accuracy (%) and classification time (s) of subset obtained by various training set reduction techniques. 117

Table 4.1 Sets of k - nearest centroid neighbors for each training sample of the classes c_1 and c_2 before the introduction of the test 144 sample, x in an offline stage (k = 3).

Table 4.2 New sets of k - nearest neighbors for each training sample of the classes c_1 and c_2 after the introduction of the test sample, x in an online stage (k = 3). (Assume that x belongs to the class c_1). 146

Table 4.3 New sets of k - nearest neighbors for each training sample of the classes c_1 and c_2 the introduction of the test sample, x in an online stage (k = 3). (Assume that x belongs to the class c_2). 147

Table 5.1 List of the Real-world selected data sets. 170

Table 5.2 Average classification accuracy (%) obtained by the RSkNCN.v2 on 30 sets of the Real-world data by varying the parameter, α for k = 3, 5, 7. 176

Table 5.3 Average classification accuracy (%) obtained by the RSkNCN.v4 on 30 sets of the Real-world data by varying the parameter, α for k = 3, 5, 7. 176

Table 5.4 Classification accuracy (%) of the RSkNCN.v2 on the FV-USM image database for when varying parameter, α for k = 1,3,5,7,9. 179

Table 5.5 Classification accuracy (%) of the RSkNCN.v4 on the FV-USM image database when varying the parameter, α for k = 1,3,5,7,9. 179

Table 5.6 Average classification accuracy (%) obtained by the kNCN-based training set reduction techniques on 30 sets of the Real-world data for k = 3, 5, 7. 182

Table 5.7 Average reduction ratio and ranking obtained by the kNCN-based training set reduction techniques on 30 sets of the Real-world data for k = 3, 5, 7. 184

Table 5.8 Average learning time (s) obtained by the kNCN-based training set reduction techniques on 30 sets of the Real-world data for k = 3, 5, 7. 187

Table 5.9 Average classification time (s) and ranking obtained by the kNCN-based training set reduction techniques on 30 sets of the Real-world data for k = 3, 5, 7. 188

Table 5.10 Learning times (s) of the kNCN-based training set reduction techniques on the FV-USM image database for k = 1, 3, 5, 7, 9. 193

Table 5.11 Average classification time (s) of the ENCN with training set reduction techniques on 30 sets of the Real-world data for k=3, 5, 7, 9. 203

Table 5.12 Classification comparison of the RSENCN and competing classifiers in terms of classification accuracy (%) with the 212 standard deviation and values of *k* in the parentheses on 30 sets of the Real-world data.

Table 5.13 Performance comparison of the RSENCN and competing classifiers by using the Wilcoxon Signed-Ranks test. 216

Table 5.14 The rank of classifiers by using the Friedman test. 217

Table 5.15 Results obtained by using the Holm post hoc test. 219

Table 5.16 The optimum number of features for each classifier. 227

LIST OF FIGURES

Page

Figure 1.1 k - nearest centroid neighbors distribution of the test sample, x (k = 5). 9

Figure 1.2 The situations when the test sample, x changes the k nearest centroid neighbors of 4 training samples, $(y_{95}\ , y_{154}, y_{155}, y_{76})$ when k = 5. 11

(a) $NCN_k(y_{95}, T) = \{y_{12}, y_{155}, y_{90}, y_{76}, y_{144}\}$.

(b) $NCN'_k(y_{95}, \{T + \boldsymbol{x}\}) = \{\boldsymbol{x}, y_{12}, y_{155}, y_{90}, y_{154}\}$.

(c) $NCN_k(y_{154}, T) = \{y_{57}, y_{95}, y_{107}, y_{12}, y_{151}\}$.

(d) $NCN'_k(y_{154}, \{T + \boldsymbol{x}\}) = \{y_{57}, \boldsymbol{x}, y_{107}, y_{95}, y_{71}\}$.

(e) $NCN_k(y_{155}, T) = \{y_{76}, y_{91}, y_{136}, y_{95}, y_{72}\}$.

(f) $NCN'_k(y_{155}, \{T + \boldsymbol{x}\}) = \{y_{76}, y_{91}, y_{136}, \boldsymbol{x}, y_{72}\}$.

(g) $NCN_k(y_{76}, T) = \{y_{155}, y_{119}, y_{91}, y_{72}, y_{95}\}$.

(h) $NCN'_k(y_{76}, \{T + \boldsymbol{x}\}) = \{y_{155}, y_{119}, y_{91}, y_{72}, \boldsymbol{x}\}$.

Figure 2.1 The workflow of the kNN classifier. 20

Figure 2.2 Two-class classification problem with the kNN classifier. 21

Figure 2.3 The set of k - nearest neighbors of the test sample, x and training samples ($y_{95}, y_{154,} y_{12}$, y_{155} and y_{76}) when k = 5. 27

(a) $NN_k(x, T) = \{y_{95}, y_{154}, y_{12,} y_{155}, y_{76}\}$.

(b) $NN_k(y_{95}, T) = \{x, y_{12}, y_{154}, y_{144}, y_{90}\}$.

(c) $NN_k(y_{154}, T) = \{y_{57}, x, y_{107}, y_{95}, y_{151}\}$.

(d) $NN_k(y_{12}, T) = \{y_{90}, y_{114}, y_{95}, y_{10}, y_{144}\}$.

(e) $NN_k(y_{155}, T) = \{y_{76}, y_{91}, y_{136}, y_{72}, y_{119}\}$.

(f) $NN_k(y_{76}, T) = \{y_{155}, y_{91}, y_{72}, y_{119}, y_{136}\}$.

Figure 2.4 The workflow of the ENN classifier during an (a) offline stage (without the test sample) and (b) online stage (with the test sample). 34

Figure 2.5 Distribution of training samples for a two-class classification problem without the test sample, *x*. 36

Figure 2.6 Distribution of training samples for a two-class classification problem with the introduction of the test sample, *x*. 37

Figure 2.7 The workflow of the kNCN classifier. 44

Figure 2.8 Steps of finding k nearest centroid neighbors of the test sample, x when k = 5. (a) Distribution of training and test samples for a two-class classification problem. 45

(b) $NCN_1(x, T) = \{y_{95}\}$. (c) $NCN_2(x, T) = \{y_{95}, y_{154}\}$.

(d) $NCN_3(x, T) = \{y_{95}, y_{154}, y_{29}\}$.

(e) $NCN_4(x, T) = \{y_{95}, y_{154}, y_{29}, y_{107}\}$.

(f) $NCN_5(x,T) = \{y_{95}, y_{154}, y_{29}, y_{107}, y_{144}\}$.

Figure 2.9 (a) 22 training samples of the Wdbac data set. (b) An edited subset, *S* by using the Wilson's editing kNCN technique. 54

Figure 2.10 Edited subset by using the Iterative kNCN editing technique. 57

Figure 2.11 (a) The Euclidean distance of k - nearest centroid neighbors for each training sample in Figure 2.9 (a). (b) Ranks of k - nearest centroid neighbors for each training sample according to the Euclidean distance. (c) The maximum rank of each training sample. (d) The largest rank, m defined by the Limited kNCN.v1. 62

Figure 2.12 Edited subset by using the Limited-kNCN.v1 technique. 64

Figure 2.13 (a) The Euclidean distance of k - nearest centroid neighbors for each training sample in Figure 2.9 (a). (b) Ranks of k - nearest centroid neighbors for each training sample according to the Euclidean distance. (c) The sorted maximum rank of each training sample. 65

Figure 2.14 Edited subset by using the Limited-kNCN.v2 technique. 66

Figure 3.1 The workflow of the RSkNCN.v1 during an (a) offline stage (without the test sample) and (b) online stage (with the test sample). 73

Figure 3.2 The edited subset, *S* using the RSkNCN.v1 technique 80

Figure 3.3 (a) The Euclidean distance of *k* - nearest centroid neighbors for the training set, *T* in Figure 2.9 (a). b) The ranks $of\ k$ - nearest centroid neighbors of training sample according to the Euclidean distance. (c) The maximum rank of training samples that agrees with the majority of its *k* - nearest centroid neighbors. 81

Figure 3.4 The workflow of the RSkNCN.v2 during an (a) offline stage (without the test sample) and (b) online stage (with the test sample). 83

Figure 3.5 Edited subset by using RSkNCN.v2 the technique. 88

Figure 3.6 (a) The Euclidean distance of *k* - nearest centroid neighbors for the training samples in Figure 2.9 (a). (b) The ranks $of\ k$ - nearest centroid neighbors of training samples according to the Euclidean distance. (c) The optimum rank of training samples that agrees with the majority of its *k* - nearest centroid neighbors. 89

Figure 3.7 The workflow of the RSkNCN.v3 during an (a) offline stage (without the test sample) and (b) online stage (with the test 92 sample).

Figure 3.8 Edited subset by using the RSkNCN.v3 technique. Subset to find (a) the first nearest centroid neighbor, (b) the second nearest centroid neighbor and (c) the third nearest centroid neighbor. 96

Figure 3.9 (a) The Euclidean distance of *k* - nearest centroid neighbors for the training samples in Figure 2.9 (a). b) The ranks of *k* - nearest centroid neighbors of training samples according to the Euclidean distance. (c) The maximum rank of *j*-th nearest centroid neighbor. 97

Figure 3.10 The workflow of the RSkNCN.v4 during an (a) offline stage (without the test sample) and (b) online stage (with the test sample). 99

Figure 3.11 Edited subset by using the RSkNCN.v4 technique. Subset to find (a) the first nearest centroid neighbor, (b) the second nearest centroid neighbor and (c) the third nearest centroid neighbor. 104

Figure 3.12 (a) The ranks of *k* - nearest centroid neighbors for the training samples in Figure 2.9 (a). (b) The sorted ranks for each *j*-th nearest centroid neighbor of training samples. 105

Figure 3.13 Training samples distribution of the Wdbc data set. 106

Figure 3.14 Subset edited by using the Wilson’s Edited kNCN. 107

Figure 3.15 Subset edited by using the Iterative kNCN. 107

Figure 3.16 Subset edited by using the (a) Limited-kNCN.v1 and (b) Limited-kNCN.v2. 109

Figure 3.17 (a) Atypical samples (outside of the circles) in the subset edited with the Limited-kNCN.v1. (b) The enlarged figure of the sample distribution in Figure 3.17 (a). (c) Subset edited with the RSkNCN.v1. 110

Figure 3.18 (a) Atypical samples (outside of the circles) in the subset edited with the RSkNCN.v1. (b) Subset edited with the RSkNCN.v2. 112

Figure 3.19 Different subsets obtained by using the RSkNCN.v3 technique to find *k* - nearest centroid neighbors when *k* = 3. (a) Subset to find the first nearest centroid neighbor. (b) Subset to find the second nearest centroid neighbor. (c) Subset to find the third nearest centroid neighbor. 114

Figure 3.20 Different subsets obtained by using the RSkNCN.v4 technique when *k* = 3. (a) Subset to find the first nearest centroid neighbor. (b) Subset to find the second nearest centroid neighbor. (c) Subset to find the third nearest centroid neighbor. 115

Figure 3.21 k - nearest centroid neighbors of the test sample, x when k = 3. $NCN_k(x) = \{y_{436}, y_{152}, y_{209}\}, C_k(x) = \{c_1, c_2, c_1\}$. 118

Figure 3.22 (a) Black circles denote the samples removed by using the Wilson's Edited kNCN. (b) k - nearest centroid neighbors of the test sample, x on the subset obtained by using the Wilson's Edited kNCN when k = 3. $NN_k(x, S) = \{y_{326}, y_{319}, y_{127}\}$, $C_k(x, S) = \{c_2, c_2, c_2\}$. (c) Black circles denote the samples removed by using the Iterative kNCN. (d) k - nearest centroid neighbors of the test sample, x on the subset obtained by using the Iterative kNCN when k = 3. $NN_k(x, S) = \{y_{278}, y_{178}, y_{260}\}$, $C_k(x, S) = \{c_2, c_2, c_1\}$. 120

Figure 3.23 k - nearest centroid neighbors of the test sample, x on the subset, S obtained by using the (a) Limited-kNCN.v1 and (b) Limited-kNCN.v2 when k = 3, $NCN_k(x, S) = \{y_{436}, y_{152}, y_{209}\}$, $C_k(x, S) = \{c_1, c_2, c_1\}$. 121

Figure 3.24 k - nearest centroid neighbors of the test sample, x on the subset, S obtained by using the (a) RSkNCN.v1 and (b) RSkNCN.v2 when $\underline{k}$ = 3, $NCN_k(x, S) = \{y_{436}, y_{152}, y_{209}\}$, $C_k(x, S) = \{c_1, c_2, c_1\}$. 121

Figure 3.25 k - nearest centroid neighbors of the test sample, x on k subsets obtained by using the RSkNCN.v3 when k = 3. First nearest centroid neighbor is found on the first subset, S_1 (a). Second nearest centroid neighbor is found on the second subset, S_2 (b) and third nearest centroid neighbor is found on the third subset, S_3 (c). $NCN_k(x, S_k) = \{y_{436}, y_{152}, y_{209}\}, C_k(x, S_k) = \{c_1, c_2, c_1\}$ 123

Figure 3.26 k - nearest centroid neighbors of the test sample, x on k subsets obtained by using the RSkNCN.v4 when k = 3. First nearest centroid neighbor is found on the first subset, S_1 (a). Second nearest centroid neighbor is found on the second subset, S_2 (b) and the third nearest centroid neighbor is found on the third subset, S_3 (c). $NCN_k(x, S_k) = \{y_{436}, y_{152}, y_{209}\}$ $C_k(x, S_k) = \{c_1, c_2, c_1\}$ 124

Figure 4.1 The workflow of the ENCN classifier during an (a) offline stage (without the test sample) and (b) online stage (with the test sample). 131

Figure 4.2 Distribution of k - nearest centroid neighbors of the test sample, x when k = 3. 141

Figure 4.3 Distribution of training samples for a two-class classification problem without the test sample, x. 143

Figure 4.4 Changes of k - nearest centroid neighbors for 3 training samples, (y_4, y_7, y_{16}) when k = 3 with the introduction of the test sample, x . (a-i), (b-i),(c-i) k - nearest centroid neighbors of 145 training samples, (y_4, y_7, y_{16}) without the test sample, x. (a-ii), (b-ii),(c-ii) k - nearest centroid neighbors of training samples, (y_4, y_7, y_{16}) with the test sample, x.

Figure 4.5 The workflow of the RSENCN classifier during an (a) offline stage (without the test sample) and (b) online stage (with the test sample). 152

Figure 4.6 (a) Ranks of k - nearest centroid neighbors for each training sample in Table 4.1 according to the Euclidean distance. (b) Ranks of selected training samples that agree with the majority of its k - nearest centroid neighbors. (c) The maximum rank of each selected training sample. 158

Figure 4.7 (a) Atypical samples (outside of the circles) in the subset edited with the RSkNCN.v1. (b) *k* nearest centroid neighbors of the test sample found in the subset. 159

Figure 4.8 (a-i), (b-i), (c-i) Subsets obtained by using the RSkNCN.v1 for training samples $\{y_4, y_7, y_{16}\}$ with atypical samples (outside of the circles). (a-ii), (b-i)i, (c-ii) The resulting *k* - nearest centroid neighbors of training samples $\{y_4, y_7, y_{16}\}$ found in the edited subset. 161

Figure 5.1 Distribution of the I-4I data set. 171

Figure 5.2 Region of interest (ROI) of finger vein image. 173

Figure 5.3 Comparison of the average reduction ratio on 30 sets of the Real-world data for different values of neighborhood, *k*. 185

Figure 5.4 Average reduction vs average accuracy (%) graph on 30 sets of the Real-world data for different values of *k*. (a) *k* = 3. (b) *k* = 5. c) *k* = 7. 186

Figure 5.5 Comparison of the average classification time (s) on 30 sets of the Real-world data for different values of neighborhood, *k*. 189

Figure 5.6 Average classification time (s) vs average accuracy (%) graph on 30 sets of the Real-world data for different values of *k*. (a) *k* = 3. (b) *k* = 5. (c) *k* = 7. 190

Figure 5.7 Classification accuracy (%) performance via different values of *k* on the FV-USM image database. 192

Figure 5.8 Classification time (s) performance via different values of *k* on the FV-USM image database. 194

Figure 5.9 Reduction ratio performance via different values of *k* on the FV-USM image database. 195

Figure 5.10 Classification accuracy (%) via different values of k on 30 sets of the Real-world data. 199

Figure 5.11 Average set size via different values of neighborhood, k on 30 sets of the Real-world data. 200

Figure 5.12 Average classification accuracy (%) vs average set size on 30 sets of Real-world data for different values of *k*. (a) *k* = 3. (b) *k* = 5. (c) *k* = 7. (d) *k* = 9. 202

Figure 5.13 Average classification accuracy (%) vs average time (s) graph on 30 sets of Real-world data for different values of *k*. (a) *k* = 3. (b) *k* = 5. (c) *k* = 7. (d) *k* = 9. 205

Figure 5.14 Classification accuracy (%) via different values of k on the FV-USM image database. 207

Figure 5.15 Comparison of the set size for different values of neighborhood, *k* on the FV-USM image database. 208

Figure 5.16 Comparison of the classification time (s) for different values of neighborhood, *k* on the FV-USM image database. 209

Figure 5.17 Performance comparison of different classifiers on 30 sets of the Real-world data using the Bonferroni-Dunn test. All classifiers that are significantly different at $\alpha_s = 0.05$ have ranking outside of the marked interval. 220

Figure 5.18 Classification accuracies (%) of kNN, kNCN, DWkNCN, ENN, DWkNN, FkNN, kGNN, MkNN and RSENCN classifiers on the I-4I data set for different values of neighborhood size, *k*. 222

Figure 5.19 Classification accuracies (%) of the kNN, kNCN, DWkNCN, ENN, DWkNN, FkNN, kGNN, MkNN and RSENCN classifiers on the I-4I data set for different values of training samples size, *N*.(a) *k* = 7. (b) *k* = 9. 223

Figure 5.20 Classification accuracies (%) of the kNN, kNCN, DWkNCN, ENN, DWkNN, FkNN, kGNN, MkNN and RSENCN classifiers on the I-4I data set for different values of noise. (a) *k* = 7. (b) *k* =9. 224

Figure 5.21 Classification accuracies (%) of the kNN, kNCN, DWkNCN, ENN, DWkNN, FkNN, kGNN, MkNN and RSENCN classifiers on the FV-USM image database by varying the number of features. 225

Figure 5.22 Classification accuracies (%) of the kNN, kNCN, DWkNCN, ENN, DWkNN, FkNN, kGNN, MkNN and RSENCN classifiers on the FV-USM image database by varying the neighborhood size, *k*. 237

LIST OF ABBREVIATIONS

CA	Classification Accuracy
CD	Critical Diagram
DWkNN	Distance Weighted *k* - Nearest Neighbor
ENCN	Extended Nearest Centroid Neighbor
ENN	Extended Nearest Neighbor
FkNCN	Fuzzy *k* - Nearest Centroid Neighbor
FkNN	Fuzzy *k* - Nearest Neighbor
FV-USM	Finger Vein Universiti Sains Malaysia
HWkNCN	Heated Weight *k* - Nearest Centroid Neighbor
IBG	Intelligent Biometric Group
KEEL	Knowledge Extraction Evolutionary Learning
kGNN	*k* - General Nearest Neighbor
kNCN	*k* - Nearest Centroid Neighbor
kNN	*k* - Nearest Neighbor
LMkNCN	Local Mean *k* - Nearest Centroid Neighbor
MkNN	Mutual *k* - Nearest Neighbor
NCN	Nearest Centroid Neighbor
NN	Nearest Neighbor
PCA	Principle Component Analysis
PNCN	Pseudo Nearest Centroid Neighbor
ROI	Region of Interest
RSENCN	Reduced Set Extended Nearest Centroid Neighbor
RSkNCN.v1	Reduced Set *k* - Nearest Centroid Neighbor.v1

RSkNCN.v2	Reduced Set *k* - Nearest Centroid Neighbor.v2
RSkNCN.v3	Reduced Set *k* - Nearest Centroid Neighbor.v3
RSkNCN.v4	Reduced Set *k* - Nearest Centroid Neighbor.v4
SE	Standard Error
UCI	University of California Irvine
USM	Universiti Sains Malaysia
UWkNCN	Uniform Weight *k* - Nearest Centroid Neighbor
WkNCN	Weighted *k* - Nearest Centroid Neighbor

LIST OF SYMBOLS

$\delta\left(c_j = c_x^i\right)$	Kronecker delta function
μ_i	Mean vector
α	Percentage of training samples with atypical samples as one of its *k* - nearest centroid neighbors
α_s	Statistical significance level
α_{adj}	The adjusted statistical significance level
μ_i	Mean vector
Σ_i	Covariance
c_x	Class of the test sample, *x*
c_i	Class *i*
d_j	The distance of the j-th nearest neighbor
$d(x, y_i)$	Distance between the test sample and training sample
f_{v1}	The fraction of training set by using the Limited-kNCN.v1
f_{v2}	The fraction of training set by using the Limited-kNCN.v2
$f_{subset.v1}$	The fraction of training set by using the RSkNCN.v1
$f_{subset.v2}$	The fraction of training set by using the RSkNCN.v2
$f_{subset.v3}$	The fraction of training set by using the RSkNCN.v3
$f_{subset.v4}$	The fraction of training set by using the RSkNCN.v4
$f_{j,v4}$	The fraction of training set for *j*-th nearest centroid neighbor by using the RSkNCN.v4
$f_{j,v3}$	The fraction of training set for *j*-th nearest centroid neighbor by using the RSkNCN.v3
f_{ENN,c_i}	Target function of ENN for class c_i
f_{ENCN,c_i}	Target function of ENCN for class c_i

$I_r(y_i, T)$	Indicator functions
k	Size of the neighborhood
k_{NN,c_i}	Number of nearest neighbors of the test sample, x from class c_i
k_{NCN,c_i}	Number of nearest centroid neighbors of the test sample, x from class c_i
M_{opt}	Optimum rank
$M_{opt,j}$	Optimum rank, $M_{opt,j}$ for j-th nearest centroid neighbor
$M_{rank,j}$	Maximum rank of j-th nearest centroid neighbor
M_{rank}	The largest rank
m	Number of classes
m_w	A parameter to determine the weighted of the distance between the test sample, x and k - nearest neighbors
m_{robust}	Robust rank
$m_{max,i}$	Maximum rank for the training sample, y_i
N	Number of training samples
$N^{\delta x}(x)$	Neighborhood information of the test sample, x
$NCN_k(x, T)$	Set of k - nearest centroid neighbors of the test sample, x
$NCN'_k(x, T)$	A new set of the k - nearest centroid neighbors of the test sample, x
$NN_k(x, T)$	Set of k - nearest neighbors of the test sample, x
$NN'_k(x, T)$	A new set of k - nearest neighbors of the test sample, x
ncn_x^r	r –th centroid neighbor of the test sample, x
n_2	Number of data sets
n_1	Number of classifiers
n_{c_i}	Total number of training samples from the same class

$\Delta n_{c_i}^{c_j}$	Number of number of training samples of class c_i which has an increased (decreased) number of samples from class c_i in its *k* - nearest neighbors or *k* centroid neighbors
p_n	Non-parametric density estimation
P	Statistical significant value
q	Total number of the targeted classes
$\bar{R}$	Mean rank
R^-	Negatives rank
R^+	Positives rank
R_i	Set of the rank for the training sample, y_i
R^p	Feature space
$r_{nn,j}$	Rank of j-th nearest centroid neighbor
S_k	Total test samples in the *k*-th fold
S	Subset
S_{c_i}	Subset of training samples from class c_i
$T_{NCN}{}_{c_i}^{c_j}$	Generalized class-wise statistic of ENCN for class, c_i when the test sample, *x* is assumed to belong to the class, c_j
$T_{NN}{}_{c_i}^{c_j}$	Generalized class-wise statistic for ENN for class, c_i when the test sample, *x* is assumed to belong to the class, c_j
TP_k	Number of true positives the *k*-th fold
TN_k	Number of true negatives the *k*-th fold
T	Training set
t	Width of the Heat Kernel function
T_{NCN,c_i}	Generalized class-wise statistic of ENCN for class, c_i
T_{NN,c_i}	Generalized class-wise statistic of ENN for class, c_i
U_j	Membership of j-th nearest neighbor

V_j	Vote of the j-th nearest neighbor
V	Volume that contains k training samples
w_j	Weight, w_j for a j-th nearest neighbor
w_i^{NCN}	Weight of j-th nearest centroid neighbors
X_F^2	Friedman statistics
x	Test sample
y_{ri}^c	Centroid point
y_i	i-th training sample
Z	Statistical significant value

TEKNIK LANJUTAN JIRAN SENTROID TERDEKAT DENGAN PENGURANGAN SET LATIHAN UNTUK PENGELASAN

ABSTRAK

Jiran Sentroid *k* Terdekat (kNCN) adalah pengelas bukan parametrik yang terkenal yang menunjukkan prestasi yang luar biasa dalam pengelasan. Namun begitu, teknik ini mempunyai masalah daripada segi masa pengelasan yang perlahan dan pemilihan satu sisi jiran sentroid terdekat yang membawa kepada prestasi ketepatan pengelasan yang lemah. Tesis ini membentangkan empat varian teknik pengurangan set data latihan yang dipanggil Pengurangan Jiran Sentroid *k* Terdekat.v1 (RSkNCN.v1), Pengurangan Jiran Sentroid *k* Terdekat.v2 (RSkNCN.v2), Pengurangan Jiran Sentroid *k* Terdekat.v3 (RSkNCN.v3) dan Pengurangan Jiran Sentroid *k* Terdekat.v4 (RSkNCN.v4) dicadangkan untuk mengurangkan masa pengelasan kNCN. Sampel atipikal dikeluarkan terlebih dahulu dengan menggunakan teknik Edit Wilson dan pecahan set latihan ditentukan menggunakan pangkat maksimum atau optimum sampel latihan (yang bersetuju dengan majoriti jiran sentroid *k* terdekatnya). Hasil eksperimen yang dijalankan ke atas tiga puluh data dunia-nyata dan data imej FV-USM menunjukkan semua teknik pengurangan latihan yang dicadangkan mencapai prestasi terbaik daripada segi nisbah pengurangan dan masa pengelasan berbanding dengan teknik penanda aras (Wilson's Edited, Iterative and Limited-kNCNs). Semua teknik pengurangan latihan yang dicadangkan mencapai keputusan yang memuaskan daripada segi ketepatan pengelasan kecuali untuk RSkNCN.v4. Teknik ini melakukan strategi penyingkiran sampel yang agresif. Oleh itu, ada kemungkinan bahawa sampel latihan yang mempunyai maklumat yang berguna telah disingkirkan menyebabkan kepada prestasi ketepatan pengelasan yang lemah. Berkenaan dengan masalah kedua kNCN, tesis ini mencadangkan Pengurangan Set Latihan Pengelas