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

LIST OF TABLES

LIST OF FIGURES

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

$\delta(c_j = c_x^i)$	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