Learning Techniques of the Radial Kernel Classifier

by

Mount-first Yat Fung Ng

A Thesis
Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science

Department of
Electrical and Computer Engineering
The University of Manitoba
Winnipeg, Manitoba, Canada

ISBN 0-612-13402-4

Name _____

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

ELECTRONICS AND ELECTRICAL ENGINEERING
SUBJECT TERM

0544
SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS
Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION
General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS
Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY
Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES
American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES
Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES
Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES
Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES
Pure Sciences
Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences
Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY
General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

Nom _____

Dissertation Abstracts International est organisé en catégories de sujets. Veuillez s.v.p. choisir le sujet qui décrit le mieux votre thèse et inscrivez le code numérique approprié dans l'espace réservé ci-dessous.

_____ ☐☐☐ U·M·I

SUJET — CODE DE SUJET

Catégories par sujets

HUMANITÉS ET SCIENCES SOCIALES

COMMUNICATIONS ET LES ARTS
Architecture 0729
Beaux-arts 0357
Bibliothéconomie 0399
Cinéma 0900
Communication verbale 0459
Communications 0708
Danse 0378
Histoire de l'art 0377
Journalisme 0391
Musique 0413
Sciences de l'information 0723
Théâtre 0465

ÉDUCATION
Généralités 515
Administration 0514
Art 0273
Collèges communautaires 0275
Commerce 0688
Économie domestique 0278
Éducation permanente 0516
Éducation préscolaire 0518
Éducation sanitaire 0680
Enseignement agricole 0517
Enseignement bilingue et multiculturel 0282
Enseignement industriel 0521
Enseignement primaire. 0524
Enseignement professionnel 0747
Enseignement religieux 0527
Enseignement secondaire 0533
Enseignement spécial 0529
Enseignement supérieur 0745
Évaluation 0288
Finances 0277
Formation des enseignants 0530
Histoire de l'éducation 0520
Langues et littérature 0279
Lecture 0535
Mathématiques 0280
Musique 0522
Orientation et consultation 0519
Philosophie de l'éducation 0998
Physique 0523
Programmes d'études et enseignement 0727
Psychologie 0525
Sciences 0714
Sciences sociales 0534
Sociologie de l'éducation 0340
Technologie 0710

LANGUE, LITTÉRATURE ET LINGUISTIQUE
Langues
- Généralités 0679
- Anciennes 0289
- Linguistique 0290
- Modernes 0291

Littérature
- Généralités 0401
- Anciennes 0294
- Comparée 0295
- Médiévale 0297
- Moderne 0298
- Africaine 0316
- Américaine 0591
- Anglaise 0593
- Asiatique 0305
- Canadienne (Anglaise) 0352
- Canadienne (Française) 0355
- Germanique 0311
- Latino-américaine 0312
- Moyen-orientale 0315
- Romane 0313
- Slave et est-européenne 0314

PHILOSOPHIE, RELIGION ET THÉOLOGIE
Philosophie 0422
Religion
- Généralités 0318
- Clergé 0319
- Études bibliques 0321
- Histoire des religions 0320
- Philosophie de la religion 0322

Théologie 0469

SCIENCES SOCIALES
Anthropologie
- Archéologie 0324
- Culturelle 0326
- Physique 0327

Droit 0398
Économie
- Généralités 0501
- Commerce-Affaires 0505
- Économie agricole 0503
- Économie du travail 0510
- Finances 0508
- Histoire 0509
- Théorie 0511

Études américaines 0323
Études canadiennes 0385
Études féministes 0453
Folklore 0358
Géographie 0366
Gérontologie 0351
Gestion des affaires
- Généralités 0310
- Administration 0454
- Banques 0770
- Comptabilité 0272
- Marketing 0338

Histoire
- Histoire générale 0578
- Ancienne 0579
- Médiévale 0581
- Moderne 0582
- Histoire des noirs 0328
- Africaine 0331
- Canadienne 0334
- États-Unis 0337
- Européenne 0335
- Moyen-orientale 0333
- Latino-américaine 0336
- Asie, Australie et Océanie 0332

Histoire des sciences 0585
Loisirs 0814
Planification urbaine et régionale 0999
Science politique
- Généralités 0615
- Administration publique 0617
- Droit et relations internationales 0616

Sociologie
- Généralités 0626
- Aide et bien-être social 0630
- Criminologie et établissements pénitentiaires 0627
- Démographie 0938
- Études de l' individu et de la famille 0628
- Études des relations interethniques et des relations raciales 0631
- Structure et développement social 0700
- Théorie et méthodes. 0344
- Travail et relations industrielles 0629

Transports 0709
Travail social 0452

SCIENCES ET INGÉNIERIE

SCIENCES BIOLOGIQUES
Agriculture
- Généralités 0473
- Agronomie. 0285
- Alimentation et technologie alimentaire 0359
- Culture 0479
- Élevage et alimentation 0475
- Exploitation des pâturages 0777
- Pathologie animale 0476
- Pathologie végétale 0480
- Physiologie végétale 0817
- Sylviculture et faune 0478
- Technologie du bois 0746

Biologie
- Généralités 0306
- Anatomie 0287
- Biologie (Statistiques) 0308
- Biologie moléculaire 0307
- Botanique 0309
- Cellule 0379
- Écologie 0329
- Entomologie 0353
- Génétique 0369
- Limnologie 0793
- Microbiologie 0410
- Neurologie 0317
- Océanographie 0416
- Physiologie 0433
- Radiation 0821
- Science vétérinaire 0778
- Zoologie 0472

Biophysique
- Généralités 0786
- Medicale 0760

SCIENCES DE LA TERRE
Biogéochimie 0425
Géochimie 0996
Géodésie 0370
Géographie physique 0368
Géologie 0372
Géophysique 0373
Hydrologie 0388
Minéralogie 0411
Océanographie physique 0415
Paléobotanique 0345
Paléoécologie 0426
Paléontologie 0418
Paléozoologie 0985
Palynologie 0427

SCIENCES DE LA SANTÉ ET DE L'ENVIRONNEMENT
Économie domestique 0386
Sciences de l'environnement 0768
Sciences de la santé
- Généralités 0566
- Administration des hôpitaux 0769
- Alimentation et nutrition 0570
- Audiologie 0300
- Chimiothérapie 0992
- Dentisterie 0567
- Développement humain 0758
- Enseignement 0350
- Immunologie 0982
- Loisirs 0575
- Médecine du travail et thérapie 0354
- Médecine et chirurgie 0564
- Obstétrique et gynécologie 0380
- Ophtalmologie 0381
- Orthophonie 0460
- Pathologie 0571
- Pharmacie 0572
- Pharmacologie 0419
- Physiothérapie 0382
- Radiologie 0574
- Santé mentale 0347
- Santé publique 0573
- Soins infirmiers 0569
- Toxicologie 0383

SCIENCES PHYSIQUES
Sciences Pures
Chimie
- Généralités 0485
- Biochimie 487
- Chimie agricole 0749
- Chimie analytique 0486
- Chimie minérale 0488
- Chimie nucléaire 0738
- Chimie organique 0490
- Chimie pharmaceutique 0491
- Physique 0494
- PolymÈres 0495
- Radiation 0754

Mathématiques 0405
Physique
- Généralités 0605
- Acoustique 0986
- Astronomie et astrophysique 0606
- Électronique et électricité 0607
- Fluides et plasma 0759
- Météorologie 0608
- Optique 0752
- Particules (Physique nucléaire) 0798
- Physique atomique 0748
- Physique de l'état solide 0611
- Physique moléculaire 0609
- Physique nucléaire 0610
- Radiation 0756

Statistiques 0463

Sciences Appliqués Et Technologie
Informatique 0984
Ingénierie
- Généralités 0537
- Agricole 0539
- Automobile 0540
- Biomédicale 0541
- Chaleur et thermodynamique 0348
- Conditionnement (Emballage) 0549
- Génie aérospatial 0538
- Génie chimique 0542
- Génie civil 0543
- Génie électronique et électrique 0544
- Génie industriel 0546
- Génie mécanique 0548
- Génie nucléaire 0552
- Ingénierie des systèmes 0790
- Mécanique navale 0547
- Métallurgie 0743
- Science des matériaux 0794
- Technique du pétrole 0765
- Technique minière 0551
- Techniques sanitaires et municipales 0554
- Technologie hydraulique 0545

Mécanique appliquée 0346
Géotechnologie 0428
Matières plastiques (Technologie) 0795
Recherche opérationnelle 0796
Textiles et tissus (Technologie) 0794

PSYCHOLOGIE
Généralités 0621
Personnalité 0625
Psychobiologie 0349
Psychologie clinique 0622
Psychologie du comportement 0384
Psychologie du développement 0620
Psychologie expérimentale 0623
Psychologie industrielle 0624
Psychologie physiologique 0989
Psychologie sociale 0451
Psychométrie 0632

LEARNING TECHNIQUES OF THE

RADIAL KERNEL CLASSIFIER

BY

MOUNT-FIRST YAT FUNG NG

A Thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

I hereby declare that I am the sole author of this thesis.

I authorize the University of Manitoba to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Mount-first Yat Fung Ng

I further authorize the University of Manitoba to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Mount-first Yat Fung Ng

Abstract

Radial Kernel Classifier (RKC) is a hybrid between the Classical Kernel Classifier (CKC) and the Radial Basis Functions Network (RBFN). RKC inherits the ability to converge to the Bayes Error from CKC and the compactness of the RBFN. In this thesis, the performance of RKC using different learning methods is compared in order to determine a proper training procedure. From the results of the experiments in this research, the following procedure is recommended for training the RKC. First, select centroids from each class separately using the K-Means clustering. Each class should have the same number of centroids. Second, set the weight of each centroid to the number of training data grouped within its cluster. Third, use the training data to estimate the covariance matrix of each class for the use of the One-Class-One-Sigma Mahalanobis metric. Finally, optimize the smoothing parameter by using the Three-Point Search method for learning h and the Leave-One-Out method for estimating the classification error. The Gaussian Kernel should be used through out the procedure.

Acknowledgments

I would like to thank Prof. Dr. Pawlak for his support and guidance throughout the course of this study. By providing me with insightful advice and explanation, he let me realize true potential behind the Radial Kernel Classifier. I would also like to express my thanks to:

Prof. Dr. Shwedyk for his advice, suggestions and for being my chairing supervisor during Prof. Dr. Pawlak's sabbatical;

Prof. Dr. Peters for being my internal examiner;

Prof. Dr. Liao for being my external examiner;

Dr. Richard Lippmann for providing me with the Vowel data in the experiments used in this thesis;

my girlfriend, Fung Yee Kwok, and my parents for their support, patient and understanding.

The research of this thesis is sponsored by MicroNet.

Contents

List of Figures

Chapter 1

Introduction

In general, classification problems can be divided into two types. In the first type, one has a set of objects with no knowledge about their class membership and it is desired to impose a class structure on them. This type of problem is called cluster analysis. In the second type of problem, a set of sample objects with known classification are available and one wishes to use them to devise a classification rule to be used for future objects. This second type of classification is called pattern recognition which is the focus of this thesis.

Pattern recognition does not necessary mean the classification of images. Speech identification, system fault diagnosis, and even diagnosis of low back disorders are pattern recognition problems.

There are two main methodologies in pattern recognition. The first is a Statistical approach while the second one is based on a the Neural Network approach. The Statistical approach consists of three types of classifiers which differ in the amount of information one knows about the data to be classified and the type of assumption made about the data.

The first type of classifier is called the Bayes classifier. This classifier can achieve the minimum classification error for any problem providing that the a priori proba-

bilities and the class-conditional densities are known for all classes. This minimum error is called the Bayes error. Since one rarely has all the information necessary to apply the Bayes classifier, it is used mainly in theoretical studies and simulations. More detail about this method is provided in Chapter 2.

When a complete statistical knowledge about the the data is not available, one can assume that the data comes from a specific form of density such as the Gaussian distribution, and proceed to estimate the necessary parameters of this density using training data. With this density function and the estimated parameters, the Bayes classifier can be used for classification. This second type of classifier is called parametric pattern classifier. A more detailed explanation of this method can be found in the book [1] by Duda and Hart.

In the cases where there is not enough knowledge to make assumptions about the underlying density of the data or when making assumptions about the density of the data is not desirable, then the third type of classifier, the non-parametric classifier is used. An example of this type of classifier is the Classical Kernel Classifier described in Chapter 2.

1.1 Problem

Although both the Statistical and the Neural Network approaches or classifiers are used to solve the same types of problems, their advantages and disadvantages are often complementary.

On one end of the spectrum are the Statistical classifiers, most of which are capable of converging to the Bayes error when the number of training data approach infinity and there are only a handful of parameters to learn. Most of these classifiers are trained with non-iterative methods, thus their learning time is relatively short compared to that of the Neural Network classifiers. However, the penalty for the convergence property and short training time is the need of these classifiers to store

and to use the whole set of training data during classification. Since most of these classifiers require the use of all training data for classification, their classification speed is usually slower than those of the Neural Network classifiers.

Neural Network classifiers have properties that are in opposition to that of Statistical classifiers. The advantages of these classifiers are their short classification time which is usually independent of the number of training data and their ability to adapt to the complexity of the problem by varying the number of parameters used. In order to achieve this flexibility, most Neural Network Classifiers use iterative methods such as gradient descent to optimize their parameters. These iterative methods do not only require a lengthy learning time to converge to a solution, but there is also no guarantee that the resulting solution is the optimal one for a given classifier. In addition to their lengthy training time, there is also the doubt that Neural Network Classifiers may not converge to the Bayes error. Although many published results have shown that Neural Network Classifiers can achieve near Bayes error in simulations, the theoretical proof for their convergence is still lacking.

1.2 Purpose

It would be ideal to have a classifier which has the convergence property and a short classification time which does not depend on the number of training data. Further it should be able to adapt to the complexity of the problem at hand and should not require the use of all the training data for classification. These ideals are the motivation behind the development of the Radial Kernel Classifier (RKC) introduced in [2]. The RKC is an approach lying between the Classical Kernel Classifier (CKC) which belongs to the Statistical approach and the Radial Basis Function Network (RBFN) which belongs to the Neural Network approach. As shown in [2], RKC is capable of converging to the Bayes error and it does not require the use of all training data for its classification. Although the RKC has the potential to be an

ideal classifier, it requires a proper procedure to realize its potential. In this thesis, in order to determine a proper training procedure, the performance of the RKC using different learning methods is compared. Since RKC is derived from CKC and RBFN, most of the learning methods used in this thesis come from the literature of these two classifiers.

1.3 Outline

The organization of this thesis is as follow. In Chapter 2, four classifiers are reviewed: the Bayes Classifier, the Classical Kernel Classifier, the Radial Basis Functions Network and the Radial Kernel Classifier. After the review, the performance of RKC using different learning techniques is researched. Chapter 3 discusses the problem of selecting the number of centroids. Different methods for selecting these centroids are investigated in Chapter 4. Next, the performance of RKC using six different distance metrics is studied in Chapter 5, followed by a study of 12 different radial kernel functions in Chapter 6. Chapter 7 compares four different methods for estimating the classification error and three different methods for learning the optimum smoothing parameter. Finally, results and findings are summarized in Chapter 8.

Note that self-learning methods for selecting the number of centroids, N were not looked at. The reason for this is because for most classification problems, the N which corresponds to the minimum classification error is usually equal to n, the number of training data. Very often, this is not the N that one looks for. Rather, in most problems one would like to select the N which gives an acceptable balance between classification accuracy and speed. Thus, it is more appropriate to select N from the classification error versus N plot than by the use of any self-learning technique.

1.4 Main Results

From the results in this thesis, the following procedure is recommended to train the RKC given a S-class classification problem with n training data:

1. Separate the training data base on their classes to prepare the data for the One Class One Net (OCON) method in step 2. The advantages of the OCON are discussed in Chapter 3.

2. Select $N^{(u)}$ centroids from the u-th class using K-Means clustering. Repeat this procedure for all S classes. The number of centroids per class should be equal for each class. Although Decision Surface Mapping (DSM) could outperform K-Means clustering in certain problems, it does not provide consistence results. Therefore, the use of K-Means clustering is recommended. The study between the performance of the K-Means and the DSM technique is in Chapter 4.

3. Set the weight, $\alpha_i^{(u)}$, of centroid $\boldsymbol{c}_i{}^{(u)}$ to the number of training data from class u which belongs to its cluster $\boldsymbol{C}_i^{(u)}$.

4. Calculate the sample covariance matrix for each class. These covariance matrices are needed for the calculation of the One Class One Sigma (OCOS) Metric. OCOS should always be used as the distance metric of RKC unless it is known that the data have a uniform distribution. In this case, the Euclidean Metric should be used instead. The performance study of using different distance metrics with RKC is in Chapter 5.

5. Finally, optimize the smoothing parameter, h, by using the Three-Point Search technique for learning h and the Leave-One-Out method for estimating the classification error.

In addition, the Gaussian Kernel should be used through out this procedure. The consistent performance of the Gaussian Kernel is shown in Chapter 6. The above

procedure should be repeated for a list of N where $N = \sum_{u=1}^{S} N^{(u)}$.

Chapter 2

Introduction to Pattern Classification

This section gives a brief introduction to four classifiers: the Bayes Classifier, the Classical Kernel Classifier, the Radial Basis Functions Network and the Radial Kernel Classifier. The Bayes Classifier is an optimal classifier which gives the minimum classification error or Bayes error when all the probability distributions involved are known. The Kernel Classifier and the Radial Basis Function Networks (RBFN) look similar, yet, their learning methods and properties are quite different. The brief review of these classifiers shall serve as a foundation for the Radial Kernel Classifier introduced in Section 2.4.

Before going into the review, the appropriate notation used in the following sections is presented. Assume that there are n training patterns, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ where each pattern $\boldsymbol{x}_i$ comes from one of the S classes. The u-th class is labeled ω_u. Each pattern $\boldsymbol{x}_i$ is a d-dimensional vector-valued random variable. The state-conditional probability density function of $\boldsymbol{x}$ is $p(\boldsymbol{x}|\omega_u)$ and the a priori probability of class ω_u is $P(\omega_u)$. The posteriori probability is denoted by $P(\omega_u|\boldsymbol{x})$ which can be computed

from $p(\boldsymbol{x}|\omega_u)$ by the Bayes rule

$$P(\omega_u|\boldsymbol{x}) = \frac{P(\omega_u)p(\boldsymbol{x}|\omega_u)}{p(\boldsymbol{x})}, \tag{2.1}$$

where

$$p(\boldsymbol{x}) = \sum_{u=1}^{S} P(\omega_u)p(\boldsymbol{x}|\omega_u). \tag{2.2}$$

2.1 Bayes Classifier

The Bayes Classifier is an optimal classifier, that is, it gives the minimum classification error rate in any problem. Given an observation $\boldsymbol{x}$, the Bayes Classifier will

$$\text{assign } \boldsymbol{x} \text{ to } \omega_u \text{ if} \qquad P(\omega_u|\boldsymbol{x}) > P(\omega_v|\boldsymbol{x}) \qquad \forall u \neq v. \tag{2.3}$$

This is called the Bayes decision rule. By substituting the Bayes rule in (2.1) into (2.3) and eliminating the scaling factor $p(\boldsymbol{x})$, the Bayes decision rule can be written as follows:

$$\text{assign } \boldsymbol{x} \text{ to } \omega_u \text{ if} \qquad P(\omega_u)p(\boldsymbol{x}|\omega_u) > P(\omega_v)p(\boldsymbol{x}|\omega_v) \qquad \forall u \neq v. \tag{2.4}$$

Unless the posteriori probabilities or the a priori and the state-conditional probabilities for all classes are known, the Bayes decision rule cannot be applied directly. These probabilities are seldom available in real life problems, therefore as a result, the Bayes Classifier is used mostly in theoretical studies and is seldom used in practice.

2.2 Classical Kernel Classifier (CKC)

To perform classification when there is no knowledge about the probability structure of the data, one can use a non-parametric classifier which uses the Bayes decision rule indirectly by replacing the a priori, $P(\omega_u)$, and the state-conditional probabilities, $p(\boldsymbol{x}|\omega_u)$, with estimates. The Classical Kernel Classifier (CKC) is an example of

a non-parametric classifier. It uses $\hat{P}(\omega_u) = n^{(u)}/n$ as an estimate of the a priori probability, where $n^{(u)}$ is the number of training data from class ω_u, and it uses the Kernel Density Estimator to estimate the state-conditional probability. Given a set of $n^{(u)}$ data, $\left\{\boldsymbol{x}_i^{(u)}\right\}_{i=1}^{n^{(u)}}$, from class ω_u, the kernel density estimate of the state-conditional probability of class ω_u is

$$\hat{p}(\boldsymbol{x}|\omega_u) = \frac{1}{n^{(u)}h^d} \sum_{i=0}^{n^{(u)}} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i^{(u)}}{h}\right) \tag{2.5}$$

where d is the dimension of $\boldsymbol{x}$, h is the smoothing parameter and K is the kernel function. To design a CKC, we need to select a kernel function, K, and to find the optimal h or an estimate of it which minimizes the classification error-rate. Usually, K is a radial symmetric, unimodal probability density function such as standard normal density function

$$K(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right) \tag{2.6}$$

where d is the dimension of $\boldsymbol{x}$. To find an estimate of the optimal h is relatively straight forward if there is enough sample data. First, the sample data are separated into training and testing data. Second, the CKC is used with the training data to classify the testing data and to try to find an estimate of the optimal h which would minimize the classification error. Other methods such as the Leave-One-Out, the Jackknife and the Bootstrap techniques can also be used to estimate h. For the details of these techniques please refer to [3] and [4].

CKC is not only easy to use, but it can also converge to the Bayes error when the number of training data approaches infinity. This convergence is a very desirable feature of CKC. It however has two major problems.

The first problem is a long classification time compared to other classifiers such as the Radial Basis Functions Network (RBFN) which is briefly reviewed in the next section. The reason for this long classification time is because CKC requires the

use all training data in its classification. In order to overcome this problem, papers, such as [5], have proposed the use of the fast Fourier Transform to speed up CKC's classification time.

The second problem of the CKC is that it requires a lot of memory to store all the training data needed for classification. To reduce the memory storage requirement of the CKC, papers, such as [6] by Fukunaga and Hayes, tried to select a subset from the training data for classification while maintaining a classification error rate close to the one achieved by using all training data. For more information on CKC, please refer to [7].

2.3 Radial Basis Function Network (RBFN)

2.3.1 Architecture

A Radial Basis Function Network (RBFN) has three layers: one input layer, one hidden layer and one output layer with each layer fully connected to the next one. The number of output nodes it has is equal to the number of classes in the classification problem. The hidden layer of the RBFN is made up of neurons. Each neuron has two parameters: a prototype or centroid and a receptive field width or bandwidth. To classify an observation, $\boldsymbol{x}$, the i-th neuron in a RBFN will calculate

$$\phi_i(\boldsymbol{x}) = \phi\left(\frac{\|\boldsymbol{x} - \boldsymbol{c}_i\|}{h}\right) \tag{2.7}$$

where $\boldsymbol{c}_i$ is centroid, h is the global scaling factor which is often set to one, $\|\ldots\|$ is the distance metric which is usually taken as the Euclidean norm and ϕ is an activation function usually taken to be a Gaussian density function of the form

$$\phi(\boldsymbol{a}_i) = \exp\left(-\frac{\boldsymbol{a}_i{}^T\boldsymbol{a}_i}{\sigma_i^2}\right) \tag{2.8}$$

where $\boldsymbol{a}_i{}^T$ is the transpose of $\boldsymbol{a}_i$ and σ_i is called the bandwidth or receptive field width. Although the global scaling factor, h, and the bandwidth, σ_i seem to serve the same

purpose, they are selected independently and in different ways. The bandwidth, σ_i, is usually set to the distance from the i-th center to its nearest centroid and is usually fixed during the optimize process of the RKC. The global scaling factor, h, is selected usually by trial and error in order to improve the performance of the RBFN. Note that in RBFN the connections between the input layer and the hidden layer have no weight. After finishing their calculation, neurons pass their results to the output nodes. The u-th output node then calculates

$$m_u(\boldsymbol{x}) = \sum_{i=0}^{N} \lambda_{ui}\, \phi_i(\boldsymbol{x}) \tag{2.9}$$

where N is the number of neurons in the hidden layer, λ_{ui} is the weight between the i-th neuron and the u-th output node and ϕ_i is the activation function of the i-th neuron. Finally, the observation $\boldsymbol{x}$ is assigned to the class that corresponds to the output node which has the highest activation.

2.3.2 Contributions

After reviewing nearly over 100 studies of RBFN, the contributions of these papers are summarized into six categories:

1. core contributions;
2. selection of the location of centers, $\boldsymbol{c}_i$;
3. selection of the number of centers, N;
4. selection of the bandwidth, σ_i, and the distance metric $\|\ldots\|$;
5. selection of the smoothing parameter, h; and
6. selection of the activation function, ϕ_i.

Core Contributions

There are three core contributions in the history of RBFN. The first contribution is by Broomhead and Lowe [8] who were the first to construct the RBFN in 1989. The second contribution is by Moody and Darken [9] who have proposed the use of K-Means clustering and P-Nearest Neighbor heuristic to learn RBFN's parameters. This learning scheme has become the standard for training RBFN. The last contribution is that of Girosi and Poggio [10]. They showed that a RBFN has the "Best Approximation" property and rederived the RBFN using the regularization theory thus demonstrating the link between the two concept.

The derivation of the RBFN by Broomhead and Lowe was based on the Radial Basis Function Interpolation (RBFI) which is a strict interpolation technique in multi-dimensional space. In general, the interpolation problem can be stated as follows: Given a set of n training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, find a function m which satisfies the interpolation condition:

$$m(\boldsymbol{x}_i) = y_i, \qquad i = 1, 2, \ldots, n. \tag{2.10}$$

In Radial Basis Function Interpolation, this function m has the form

$$m(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i \, \phi(\|\boldsymbol{x} - \boldsymbol{x}_i\|) \tag{2.11}$$

where the training data $\boldsymbol{x}_i$ is used as the center, λ_i is the weight, $\{\phi(\|\boldsymbol{x} - \boldsymbol{x}_i\|)|i = 1, 2, \ldots, n\}$ is a set of n functions known as radial-basis functions, and $\|\ldots\|$ denotes a distance function which is usually taken as Euclidean. For more detail about RBFI, please refer to [11]. In order to create the network analogy, Broomhead and Lowe [8] generalized some of the assumptions of the RBFI. In particular, they relaxed the strict interpolation nature of the RBFI by selecting a random subset of Radial Basis Function 'centers' from the training data instead of using the whole training set. The

resulting RBFN has the form

$$m(\boldsymbol{x}) = \sum_{i=1}^{N} \lambda_i \, \phi(\|\boldsymbol{x} - \boldsymbol{c}_i\|), \qquad N < n \tag{2.12}$$

where $\boldsymbol{c}_i$ is the center and N is the number of centers. Once the 'centers' are chosen, the adjustable weights of the network from the hidden-to-output layers are determined by linear least-squares optimization. In other words, if

$$(\boldsymbol{Y})_i = y_i, \quad (\Lambda)_i = \lambda_i, \quad (\Phi)_{ij} = \phi(\|\boldsymbol{x}_i - \boldsymbol{c}_j\|)$$

then the adjustable weight Λ is equal to

$$\Lambda = \Phi^{+}\boldsymbol{Y}$$

where Φ^{+} is the Moore-Penrose pseudo-inverse of Φ [12].

Moody and Darken [9] proposed a network similar to RBFN but with a different name, the Local Receptive Fields (LRF). LRF is actually a normalized version of the RBFN and has the form

$$m(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} \lambda_i \phi(\|\boldsymbol{x} - \boldsymbol{c}_i\|/\sigma_i)}{\sum_{j=1}^{N} \phi(\|\boldsymbol{x} - \boldsymbol{c}_i\|/\sigma_i)} \tag{2.13}$$

where ϕ is a Gaussian function, $\boldsymbol{c}_i$ is the centers, σ_i is the bandwidth, λ_i is the weight or amplitude and N is the number of centers. They showed that using K-Means clustering to select RBFN centers gives a better performance than the use of a random subset of the training data as centers. They also proposed the use of the "P-Nearest Neighbor" heuristic for selecting the bandwidth. Under the P-Nearest Neighbor heuristic, the bandwidth is set equal to the root mean square value $\langle \boldsymbol{x}_i \rangle_P$ of the Euclidean distances from the P nearest neighbor centers. After their paper, the K-Means clustering and the P-Nearest Neighbor heuristic have become a standard RBFN learning scheme in the RBFN literature.

The third major contribution came from Poggio and Girosi who have shown that RBFN corresponds to the solution of a class of ill-posed, inverse problems involving

the reconstruction of a function from a sparse set of training data [10, 13]. In particular, they started with the concept of regularization, and derived an approximation scheme which included RBFN as a special case. Thus, they showed the close relation between these two techniques. In addition, they also showed that RBFN has the best approximation ability, an ability which MLP does not have for the class of continuous functions defined on a subset of $\mathcal{R}^d$. (An approximation scheme has the best approximation property, if in the set $\mathcal{F}$ of approximation functions, there is one that has minimum approximating error for any function to be approximated from a given set of functions.)

Selecting the Location of Centers

When RBFN was first proposed in 1988 by Broomhead and Lowe [8], they suggested that the location of the centers could either be selected uniformly within the region of $\mathcal{R}^d$ where there is data or they could be selected as a random subset of the training data. The latter is referred to in this thesis as random centers.

In respond to many critics who suggested that the choice of centers may affect the final performance of the network, Lowe showed in [14] that nonlinear optimization of the centers' locations would not improve the generalization performance of RBFN. Beastall in [15] also showed that using Kohonen's Learning Vector Quantization (LVQ) to locate RBFN centers gave no "appreciable difference" in performance compared to the use of random centers.

On the other hand, in [16] and [9], Moody and Darken showed that using adaptive K-Means clustering for locating centers gave a better performance than random centers. Also, in [17, 18], Chen et. al showed that using Orthogonal Least Squares Learning Algorithm or Orthogonal-forward-regression could also outperform random centers.

Although these results seem contradictory, most researchers in the field of RBFN

believe that using clustering techniques to locate centers does give a better performance. As a result, K-Means clustering has become the standard method for locating RBFN's centers.

In addition, in 1993, Lay and Hwang [19], and Mak, Allen and Sexton [20] showed that RBFN has a better performance when the centers are selected by clustering data from each class, One Class One Net (OCON), compared to selecting centers from all the training data, Multi-Class One Net (MCON).

Selecting the Number of Centers

Up until 1991, the only guideline for selecting the number of centers N was that N should be less than the number of training data n and the only method available to determine N was by trial and error. Starting from 1991, the importance of selecting N finally received the attention it deserves. Two different methods for selecting the number of centers have been proposed since then.

The first method changes the problem of learning the number of centers N to the problem of learning R, the radius of each cluster. The purpose is to select a radius R which minimizes the overlapping between clusters from opposite classes. This method was proposed by Musavi et. al [21] and Lemarie [22].

The second method defines an error measure or threshold which allows one to determine whether there are enough centers for the problem or whether more centers are needed. Many papers have proposed a method of this type, for example, Bye [23], Chen et. al [18], Kadirkamanathan and Niranjan [24], Katayama et. al [25], Lee and Kil [26] and Reynolds and Tarassenko [27].

Selecting the Distance Metric and Bandwidth

When Broomhead and Lowe first constructed the RBFN, it had the form

$$m(\boldsymbol{x}) = \sum_{i=0}^{N} \lambda_i \phi(\|\boldsymbol{x} - \boldsymbol{c}_i\|). \tag{2.14}$$

There was no bandwidth parameter σ_i, and the metric $\|\ldots\|$ was taken as the Euclidean distance. The bandwidth parameter first appeared in a paper by Moody and Darken [16] in 1988. Their RBFN was of the form

$$m(\boldsymbol{x}) = \sum_{i=0}^{N} \lambda_i \phi \left(\frac{\|\boldsymbol{x} - \boldsymbol{c}_i\|}{\sigma_i} \right). \tag{2.15}$$

Although from (2.15) it is clear that the bandwidth, σ_i, is not part of the distance function, $\|\ldots\|$, if $\|\boldsymbol{x} - \boldsymbol{c}_i\|/\sigma_i$ is considered as a Mahalanobis distance of the form

$$(\boldsymbol{x} - \boldsymbol{c}_i)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{c}_i) \tag{2.16}$$

with $\Sigma_i = \sigma_i I$, where I is an Identity matrix, then equation (2.15) has the same form as (2.14). If the Gaussian density function is used as the activation function, ϕ, then one could also consider the bandwidth as the variance of the Gaussian function. Since the bandwidth can have more than one interpretation, in order to be consistent, in this thesis, the bandwidth is considered as a part of the metric. With this interpretation, the Euclidean metric can be considered as a special case of the Mahalanobis metric with $\Sigma_i = I$. As a result, the only metric ever used in the field of RBFN is the Mahalanobis distance.

Although the Mahalanobis distance is the only metric used in RBFN, many methods had been proposed for selecting the bandwidth, starting with Moody and Darken who in 1968 proposed the use of P-Nearest Neighbor heuristic as the bandwidth [16].

In 1989, Houselander and Taylor [28] used a modified delta rule to learn the bandwidth. They also compared the performance of RBFN using ellipsoids with the use of spheres. In other words, they compared a RBFN using

$$\Sigma_i = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d \end{bmatrix}$$

with a RBFN using $\Sigma_i = \sigma_i\, I$. Their results showed that the RBFN using ellipsoids gave a better performance than one using spheres.

In the same year, Lowe [14] compared a RBFN using a fixed bandwidth

$$\Sigma_i = \frac{\alpha}{N} I$$

where α is the maximum distance between the chosen centers and N is the number of centers, with a RBFN using Σ_i which was learnt using a nonlinear optimization technique. His conclusion was that using a nonlinear optimization technique to learn the bandwidth did not improve the generalization capability of the RBFN.

Finally, Cios et. al [29], 1991, proposed that the bandwidth be learned by first initializing it using the P-Nearest Neighbor heuristic. Then the bandwidth is adjusted to reduce the inter-class interference between the outputs of the first layer nodes. This inter-class interference occurs when a training vector belonging to class u causes any node belonging to class v ($v \neq u$) to give an output larger than a certain threshold. This process of adjusting the bandwidth continues until no interference is present.

Selecting the Smoothing Parameter

Though the bandwidth and the smoothing parameter appear similar, the bandwidth, σ_i, is usually unique for each hidden node and it controls the spread of the radial basis function. The smoothing parameter, h, on the other hand, is a global constant or scaling factor which is used to tune the performance of the RBFN by scaling the bandwidth.

Lee was the first one to use the smoothing parameter, h, in RBFN in 1991 [30]. The proper h value was determined experimentally.

In 1991, Reynolds and Trassenko [27] proposed the use of the Locality Index method to speed up the process of estimating the smoothing parameter. Using this method, the smoothing parameter h is estimated by the value 2^ℓ where ℓ is an integer. The parameter ℓ is called the "locality index". This index method may not give the optimal smoothing parameter, nevertheless, Reynolds and Trassendko claimed that the index method is an efficient method to find an estimate of the smoothing

parameter. Although they did not develop a new method for learning the smoothing parameter, they did demonstrate the importance of the smoothing parameter by showing that there is a dependence between the RBFN's error rate and the smoothing parameter.

Selecting the Activation Function

The Gaussian function is the most used activation function in RBFN literature. It was first used in Broomhead and Lowe's paper [8] who used the Gaussian function merely as an example rather than a recommendation. At the end of their paper, they clearly stated that they had not studied which form of activation function, ϕ, should be used. Since then, no paper has studied the effect of using different activation functions such as the thin-plate spline, the multi-quadric equation and the inverse multi-quadric equation on the classification performance of RBFN. Thus far the only paper that compares the performance of seven different radial basis functions on non-linear data modeling was that by Carlin [31] in 1992. In Carlin's paper, he shows that the logarithmic function

$$\phi(x) = log(x^2 + c^2) \qquad \text{where } c \text{ is a positive constant}$$

which is a global activation function is actually better than the Gaussian function which is a local function. More experiments are necessary in order to determine whether the same result holds for classification problems.

2.4 Radial Kernel Classifier (RKC)

To conclude this chapter, the Radial Kernel Classifier (RKC) is described. It is a hybrid between the Radial Basis Functions Network and the Classical Kernel Classifiers. The classification rule of the RKC is as follows:

Given a set of S-class training data, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i$ is the i-th pattern or feature vector, y_i is its class label and n is the number of training data, the RKC will classify an observation $\boldsymbol{x}$ to class u if

$$P(\omega_u)\, q^{(u)}(\boldsymbol{x}) > P(\omega_v)\, q^{(v)}(\boldsymbol{x}), \qquad u \neq v; u, v = 1, \ldots S, \tag{2.17}$$

where

$$q^{(u)}(\boldsymbol{x}) = \sum_{i=1}^{N^{(u)}} \alpha_i^{(u)}\, \phi\left(\frac{\|\boldsymbol{x} - \boldsymbol{c}_i^{(u)}\|}{h}\right), \tag{2.18}$$

and where $P(\omega_u)$ is the a priori probability of class u, $N^{(u)}$ is the number of centroids of class u, $\boldsymbol{c}_i^{(u)}$ is the i-th centroid from class u, $\alpha_i^{(u)}$ is its weight, ϕ is the radial kernel function, $\|\cdots\|$ is the metric and h is the smoothing parameter. The remaining of this section gives a step by step derivation of the RKC.

The derivation starts by first recalling the classification rule of the Classical Kernel Classifier (CKC):

Given a set of S-class training data, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, the CKC classifies an observation $\boldsymbol{x} \in \mathcal{R}^d$ to class $u \in S$ if

$$P(\omega_u)\tilde{f}^{(u)}(x) \geq P(\omega_v)\tilde{f}^{(v)}(x), \qquad u \neq v, \tag{2.19}$$

where $\tilde{f}^{(u)}(x)$ is an estimate of the class-conditional probability density function, $p(\boldsymbol{x}|\omega_u)$ with the form

$$\tilde{f}^{(u)}(x) = \frac{1}{n^{(u)}h^d} \sum_{i=1}^{n^{(u)}} K\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}_i^{(u)}\|}{h}\right) \tag{2.20}$$

where $n^{(u)} < n$ is the number of training data from class u, K is the kernel function which is usually a distribution function, $\boldsymbol{x}_i^{(u)}$ is the i-th data from class u, $\|\cdots\|$ is the metric and h is the smoothing parameter.

In CKC, all training data are used to classify an observation. As the number of training data increases, so will the classification time of the CKC. A large number of training data is necessary because if the number is small, the CKC can not converge to the Bayes Error. In order to speed up the classification time of the CKC, equation (2.20) can be replaced with an estimate. By grouping the training data into N clusters and replacing each training data with a prototype from the cluster that it belongs to, equation (2.20) can be approximated by

$$\tilde{f}^{(u)}(x) \approx \sum_{i=1}^{N^{(u)}} \alpha_i^{(u)} \, K\left(\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}_i^{(u)}\|}{h}\right) \tag{2.21}$$

where $N^{(u)}$ is the number of clusters from class u, $\tilde{\boldsymbol{x}}_i^{(u)}$ is the i-th prototype of class u and $\alpha_i^{(u)}$ is the number of training data grouped into cluster i. Since $N < n$, using equation (2.21) to classify data should speed up the classification time. If equation (2.21) is compared with the output of the k-th output node of the RBFN

$$m_k(\boldsymbol{x}) = \sum_{i=1}^{N} \lambda_{ki} \phi\left(\frac{\|\boldsymbol{x} - \boldsymbol{c}_i\|}{h}\right), \tag{2.22}$$

the similarity in these equations can be noticed. If a class label is assigned to each RBFN centroids and λ_{ki} is set to the number of training data from class k which are grouped into the cluster i, then equation (2.21) is the same as equation (2.12). Equation (2.21) forms the basis for the Radial Kernel Classifier.

A formal definition of the Radial Kernel Classifier (RKC) is as follows:

> Given a set of S-class training data, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in \mathcal{R}^d$ is the i-th pattern or feature vector, y_i is its class label and n is the number of training data, the RKC will classify an observation $\boldsymbol{x}$ to class u if
>
> $$P(\omega_u)\, q^{(u)}(\boldsymbol{x}) \geq P(\omega_v)\, q^{(v)}(\boldsymbol{x}), \qquad u \neq v, \tag{2.23}$$

where $P(\omega_u)$ is the a priori probability and $q^{(u)}(\boldsymbol{x})$ is the estimated class-conditional probability density of class u which has the form

$$q^{(u)}(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{N^{(u)}} \alpha_i^{(u)} \phi \left(\frac{\|\boldsymbol{x} - \boldsymbol{c}_i^{(u)}\|}{h} \right), \tag{2.24}$$

where n is the number of training data, $N^{(u)}$ is the number of centroids from class u, $\boldsymbol{c}_i^{(u)}$ is i-th centroid, $\alpha_i^{(u)}$ is its weight, ϕ is the radial kernel function, $\|\ldots\|$ is the metric, h is the smoothing parameter, and d is the dimension of the vector $\boldsymbol{x}$.

The function $q^{(u)}(\boldsymbol{x})$ is called the Radial Kernel Density Estimate (RKDE). The following is an outline of the procedure for training the RKC:

1. Select N centroids from the set of n training data.
2. Assign each training data to the closest centroid to form N clusters.
3. Set $N^{(u)}$ to the number of clusters which contain data from class u.
4. Set $\alpha_i^{(u)}$ to the number of training data in cluster $\boldsymbol{C}_i$ which belongs to class u.
5. Optimize the smoothing parameter h with respect to classification error of the RKC.

The procedure described above is just an outline of how to train the RKC. In order to implement RKC, one needs to know more detail about how to learn its parameters such as how to select the location of the centroids and how many centroids should be used for a given problem. In the remaining chapters, the performance of RKC using different learning methods to train its parameters is compared. Since RKC is a hybrid between Classical Kernel Classifier (CKC) and Radial Basis Function Network (RBFN), most of the learning methods used in this study come from the CKC and RBFN literature.

To start this study, different methods used to select the number of centroids are looked at in the next chapter.

Chapter 3

Number of Centroids

In this chapter the problem of how the number of centroids should be selected for a given problem is studied. The number of centroids, N, and the location of the centroids, $\boldsymbol{c}_i$, are the two important parameters in RKC. For example, if the number of centroids used to represent the problem are sufficient, but instead of finding good locations for these centroids, they scattered randomly around in the input domain, a large classification error rate shall result even if other parameters are optimized. On the other hand, if the centroids are located optimally, but there are not enough centroids to represent our problem, the performance of the RKC will suffer. Although these two vital parameters go hand in hand, in order to have a clear understanding as to how each of these parameters affects the classification performance of the RKC, they are studied independently.

3.1 Relations with Class Memberships

The first study considers whether the class labels of the training data should be ignored when centroids are selected or whether centroids should be selected independently from each class. The former method is called the Multi-Class One Net

(MCON) and the latter method is called the One Class One Net (OCON). Since MCON uses less information in its classification, its classification error should be higher than those obtained by using OCON. The simulation results in this section show that this is in fact the case. These methods are described in detail in the next two sections.

3.1.1 Multi-Class One Net (MCON)

In MCON, the class label of the training data is ignored when the centroids, $\{\boldsymbol{c}_i\}_{i=1}^{N}$, are selected. As a result, the centroid, $\boldsymbol{c}_i$, does not have a class label. Each centroid has S weights, $\left\{\alpha_i^{(u)}\right\}_{u=1}^{S}$. Each weight, $\alpha_i^{(u)}$, corresponds to the number of training data from class u grouped into the cluster $\boldsymbol{C}_i$. Thus, using MCON the RKC will classify an observation $\boldsymbol{x}$ to the class u if

$$P(\omega_u)\sum_{i=1}^{N}\alpha_i^{(u)}\phi\left(\frac{\|\boldsymbol{x}-\boldsymbol{c}_i\|}{h}\right) \geq P(\omega_v)\sum_{i=1}^{N}\alpha_i^{(v)}\phi\left(\frac{\|\boldsymbol{x}-\boldsymbol{c}_i\|}{h}\right) \tag{3.1}$$

where $P(\omega_u)$ is the a priori probability of class u, N is the total number of centroids, ϕ is the radial kernel function, $\|\cdots\|$ is a metric, and h is the smoothing parameter.

MCON was used by Broomhead and Lowe [8] when they first proposed the RBFN in 1988. Since then almost all the RBFN papers used this method for selecting the number of centroids.

3.1.2 One Class One Net (OCON)

Unlike MCON, in OCON each centroid has a class label. Using OCON, training data are first grouped into classes, then centroids are selected from each class independently. In other words, the data in each cluster $\boldsymbol{C}_i$ can only come from one of the S classes. As a result, each centroid, $\boldsymbol{c}_i^{(u)}$, has only one weight, $\alpha_i^{(u)}$, which corresponds to the number of data from class u grouped into cluster $\boldsymbol{C}_i^{(u)}$. Under this method, each class will have its own sub-network within the RKC and hence the name One

Class One Net. Using OCON, RKC classifies an observation $\boldsymbol{x}$ to the class u if

$$P(\omega_u)\sum_{i=1}^{N^{(u)}}\alpha_i^{(u)}\phi\left(\frac{\|\boldsymbol{x}-\boldsymbol{c}_i^{(u)}\|}{h}\right) \geq P(\omega_v)\sum_{i=1}^{N^{(v)}}\alpha_i^{(v)}\phi\left(\frac{\|\boldsymbol{x}-\boldsymbol{c}_i^{(v)}\|}{h}\right) \tag{3.2}$$

where $P(\omega_u)$ is the a priori probability of class u, $N^{(u)}$ is the number of centroids that contain data from class u, ϕ is the radial kernel function, $\|\cdots\|$ is a metric, and h is the smoothing parameter.

The first paper which used OCON with RBFN was by Oglesby and Mason in 1991 [32]. In this paper, they showed that using OCON with RBFN can outperform MCON. From the RBFN papers such as [20] and [32] which advocated the use of OCON, it was not clear whether the number of centroids used for each class should always be equal, or the number of centroids should be proportional to the a priori probability of each class. In this section, equal number of centroids for each class is used. The question about the relationship between a priori probabilities and the number of centroids is studied in the next section.

To study the effect that these two methods have on the classification performance of RKC, they are compared using two experiments. The first experiment is a three classes, two dimensional classification problem. The second experiment is the vowel classification experiment described in [33] by Peterson and Barney in 1952. In both experiments, a set of training data is first generated. Secondly, these data are clustered into N groups using K-Means clustering [34]. (For more detail on K-Means clustering, please refer to Section 4.1.2.) Third, the weight, $\alpha_i^{(u)}$, is set to the number of training data from class u grouped into cluster $\boldsymbol{C}_i$. Finally, the smoothing parameter h is optimized with respect to the classification error of the RKC. Details of these experiments are in the next section.

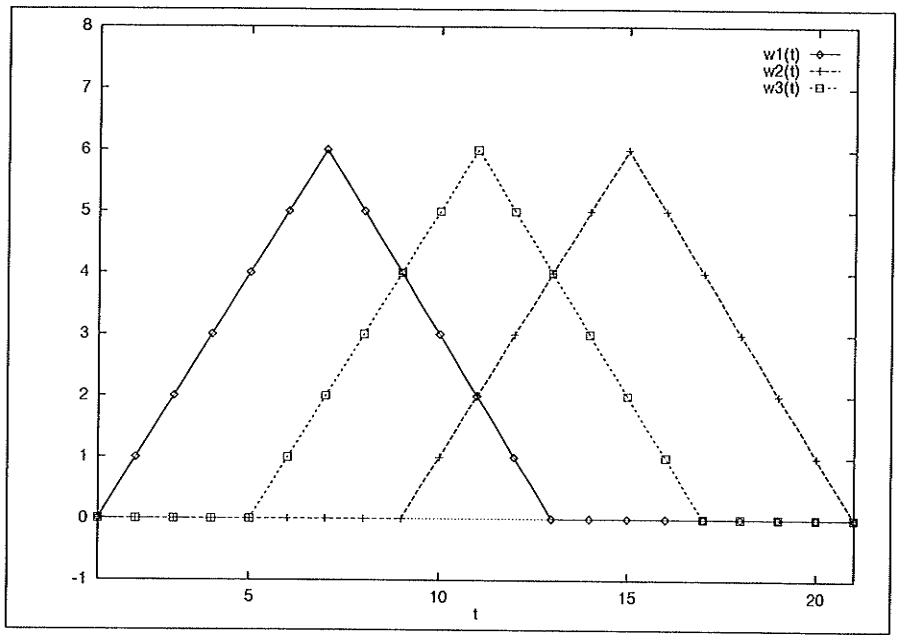

Figure 3.1: Three Waveforms, $w_1(t)$, $w_2(t)$, $w_3(t)$

3.2 Experiments

3.2.1 Waveform Classification

This experiment is a three-class, 21 dimensional waveform classification problem. This example was used in 1984 by Breiman [35].

Procedures

The waveforms used in this experiment were based on three waveforms $w_1(t)$, $w_2(t)$, $w_3(t)$ plotted in Figure 3.1. Each class consisted of a random convex combination of two of these waveforms sampled at the integer values with noise added. To generate a 21-dimensional vector, $\boldsymbol{X} = (X_1, \ldots, X_{21})$, for Class 1, a uniform random number U and 21 random numbers, $\epsilon_1, \ldots, \epsilon_{21}$, with Gaussian distribution $\mathcal{N}(0, 1)$ were

generated with

$$X_m = Uw_1(m) + (1-U)\,w_2(m) + \epsilon_m, \quad m = 1, \ldots, 21. \tag{3.3}$$

To generate a vector for Class 2, the above procedure was repeated where

$$X_m = Uw_1(m) + (1-U)\,w_3(m) + \epsilon_m, \quad m = 1, \ldots, 21. \tag{3.4}$$

A vector for Class 3 was generated by the same procedure with

$$X_m = Uw_2(m) + (1-U)\,w_3(m) + \epsilon_m, \quad m = 1, \ldots, 21. \tag{3.5}$$

For each set of training data, three hundred data were generated using a prior probabilities of $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Thus there were about 100 training data per class. Ten training data sets were generated using the above procedure together with ten test sets of size 3000.

For each set of data and a fixed number of centroids N, first the centroids are located using K-Means clustering. When OCON was used, the training data were clustered separately based on their classes, that is, it was necessary to cluster three times, once for each class, in order to locate all the centroids. When MCON was used, all the training data were clustered together, that is, only one clustering is necessary. Next, the weight, $\alpha_i^{(u)}$, was set to the number of training data grouped into cluster $\boldsymbol{C}_i$. Finally, the smoothing parameter h is optimized using a test set. In this experiment, a Gaussian distribution was used as ϕ and Euclidean metric was used as the distance metric. A list of classification errors in percentage and the corresponding h were recorded. Due to the fact that the location of the centroids which were selected using K-Means clustering depended on the location of the initial centroids, these initial centroids were selected randomly from the training data. In addition, the procedures of training the RKC were repeated ten times for each training and testing set in order to minimize this dependency. As a result, the classification errors recorded here are an average over these ten repeated training and over the ten sets of training and testing data. These results are reported in the next section.

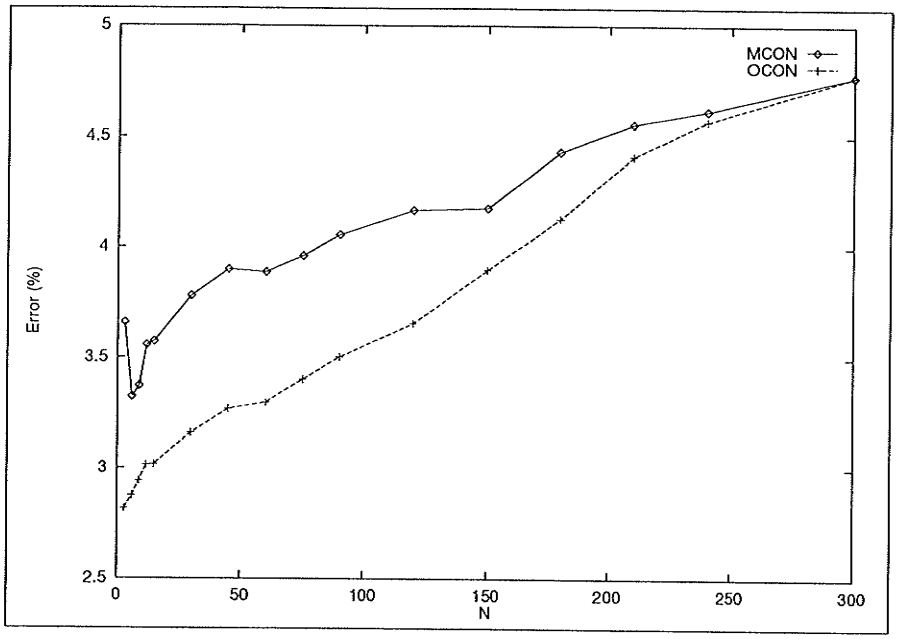

Figure 3.2: Classification results of the Waveform Experiment using Radial Kernel Classifier with Multi-Class One Net (MCON) and One Class One Net (OCON)

Results

After averaging the classification errors over ten sets of training and testing data, the error rate versus the number of centroids N curves for the MCON and the OCON is plotted in Figure 3.2. Clearly OCON outperforms MCON for every $N < n$. Also, OCON achieved a classification error smaller than the one obtained by the Classical Kernel Classifier (CKC) except for $N = n$ when the errors are equal. In this experiment, the OCON is noticed to have a faster learning time than MCON especially when N is small.

Discussion

These results show that OCON outperformed MCON for every N (except when $N = n$). Its learning time is also shorter than those of the MCON. The reason for the

shorter learning time is because MCON had to cluster all 300 data at once but OCON was allowed to cluster a small set (about 100 data) of data at a time. These results also show that OCON outperforms CKC with only three centroids even though problem is a high dimensional one. It seems that RKC did not simply reduce the number of data used in classification, it had also learned and refined the information within the training data.

3.2.2 Vowel Classification

This experiment is a ten-class, four dimensional vowel classification problem. The vowel data originated from the paper [33] by Peterson et. al. in 1952. The original data contained 76 speakers. Each speaker recorded two lists of 10 words, making a total of 1520 recorded words. The recorded words were then used to generate the four frequency variables by means of the sound spectrograph. The data set used in this paper was provided by Lippmann [36]. It contained only 75 speakers with the token [AO] of three speakers missing. As a result, the total number of data used in this experiment was 1494. A list of the vowel data used in this experiment is in Appendix A.

Procedures

About fifty training data were selected randomly without replacement from each class to make up a training set with a total of 500 data. The remaining data were then used as a test set. Each class had the same a priori probability. Ten sets of training and testing data were generated.

For each set of data and a fixed number of centroids N, first the centroids were located using K-Means clustering. When OCON was used, the training data were clustered independently based on their classes. When MCON was used, all the training data were clustered together. Next, the weight, $\alpha_i^{(u)}$, was set to the number of

training data from class u grouped into cluster $\boldsymbol{C}_i$. Lastly, the smoothing parameter h was optimized using the test set. In this experiment, a Gaussian distribution was used as ϕ and the Euclidean was used. A list of classification errors in percentage and the corresponding h were recorded. Due to the fact that the location of centroids which were selected using K-Means clustering depended on the location of initial centroids used in K-Means, these initial centroids were selected randomly from the training data. This process of training the RKC was repeated ten times for each training and testing set. As a result, the classification errors recorded here are an average over ten training processes and over the ten sets of training and testing data. These results are reported in the next section.

Results

The classification errors averaged over ten sets of training and testing data versus N curves for the MCON and the OCON are plotted in Figure 3.3. Although both MCON and OCON converged to the same minimum classification error rate when $N = n$, OCON converged faster than MCON. Note that the same minimum classification error rate could be obtained if the CKC is used in this experiment.

Discussion

From the results of this vowel experiment, OCON had outperformed MCON. Although OCON and MCON are not able to reach a smaller classification error than the one obtained by Classical Kernel Classifier, OCON had a faster convergent rate. The advantage of a having a fast convergent rate is that if one wants to reduce the classification time by using less centroids at the price of a slight increase in classification error rate then OCON as compared to MCON allows the use of a smaller N to achieve the same classification error.

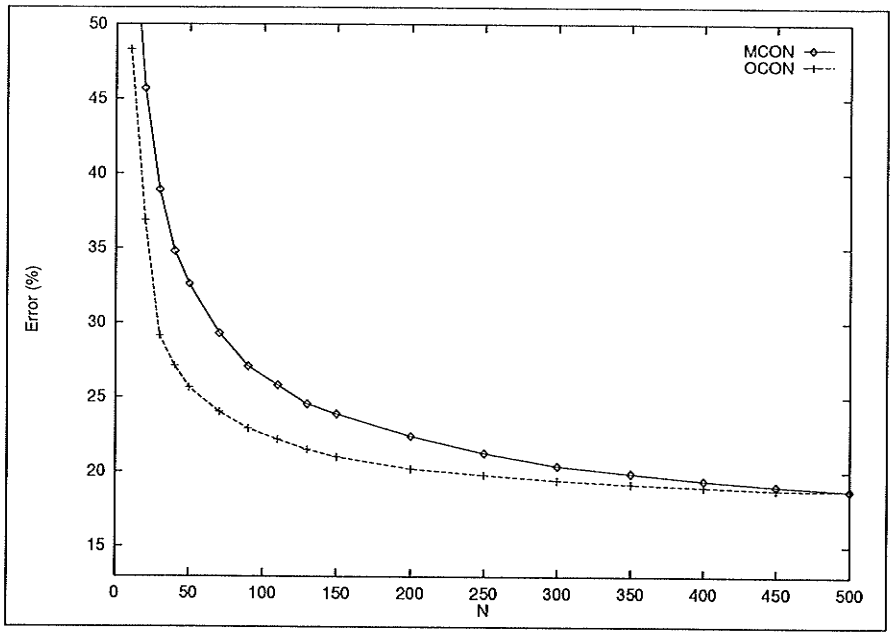

Figure 3.3: Classification results of the Vowel Experiment using the Radial Kernel Classifier with Multi-Class One Net (MCON) and One Class One Net (OCON)

3.2.3 Summary

From these two experiments, it is showed that OCON can outperform MCON. In the case where RKC could not achieve a classification error rate smaller than the one obtained by the CKC, OCON has a fast convergent rate than MCON. In addition, with OCON one can be sure that each class has at least one centroid or prototype. As a result, OCON is used in the rest of this thesis.

3.3 Relations with a Priori Probability

Now that it has been established that centroids should be selected from each class independently, the next step is to determine how many centroids should be selected from each class. An obvious method is to search through all combinations of $N^{(u)}$, where u is one of the S classes, to find the one with the minimum classification error. This method however is too time consuming even for a small number ($\approx$ 100) of training data. A second method, called the "Ratio N method" (RNM), is to let the number of centroids from each class be proportional to their a priori probabilities. A third method is to use an equal number of centroids for each class. This method is called the "Equal N method" (ENM). In this section, the performance of the RKC using RNM and ENM is studied using two examples. These two examples are basically the same except that their a priori probabilities are different.

3.3.1 Gaussian Data Classification I

This experiment is a two-class, two dimensional classification problem. Both classes have a Gaussian distribution. The first class has a zero mean with an identity covariance matrix and the second class has a mean vector $[1 \quad 2]$ and a diagonal covariance matrix with the entries 0.01 and 4. The a priori probabilities for the first and the second class are $\frac{1}{3}$ and $\frac{2}{3}$ respectively.

Procedures

Each training data set contained a total of 400 data and each test set contained 2000 data. Ten sets of training and testing data were generated. In this experiment, the OCON was used to select N; K-Means clustering for selecting centroids; Gaussian Kernel as the radial kernel function; and Euclidean distance as the metric. The initial centroids used in the K-Means clustering were selected randomly from the training data. The RKC was trained and tested ten times for each training and testing set. As a result, the classification errors recorded here are an average over these ten training and testing per data set and over the ten sets of training and testing data. The classification error rate versus N curves for the RNM and ENM methods are plotted in the next subsection.

Results

The error rate versus N curves for the RNM and the ENM are plotted in Figure 3.4. It shows that ENM has a better classification performance than RNM. These results are discussed together with the results of the next experiment in the following section.

3.3.2 Gaussian Data Classification II

This experiment is basically the same as the previous one except that the a priori probabilities are different. The a priori probabilities for the first and the second class are $\frac{2}{3}$ and $\frac{1}{3}$ respectively. The training procedures of this simulation is the same as the previous one so they are not repeated here.

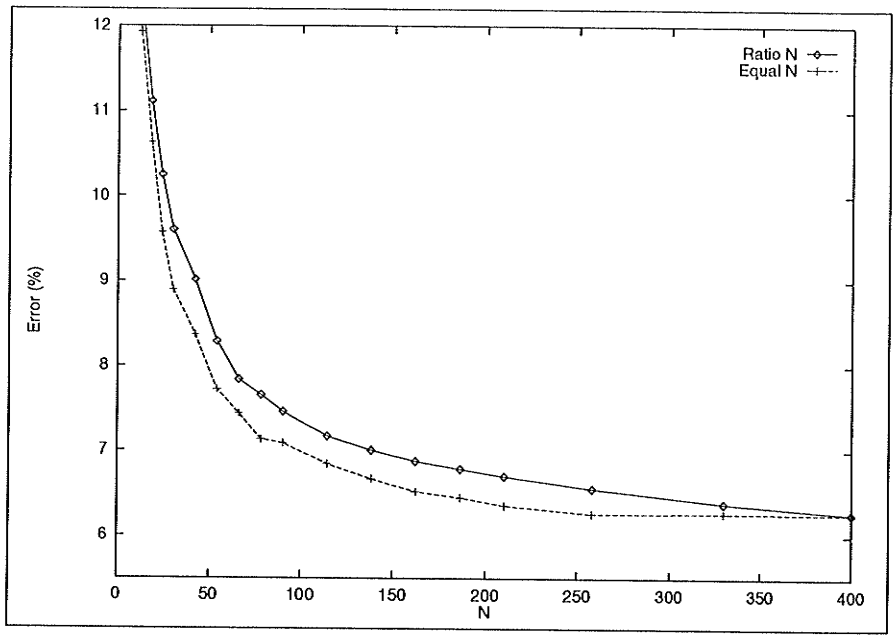

Figure 3.4: Classification results of the Gaussian Data Classification I using the Radial Kernel Classifier with the Ratio N method and the Equal N method

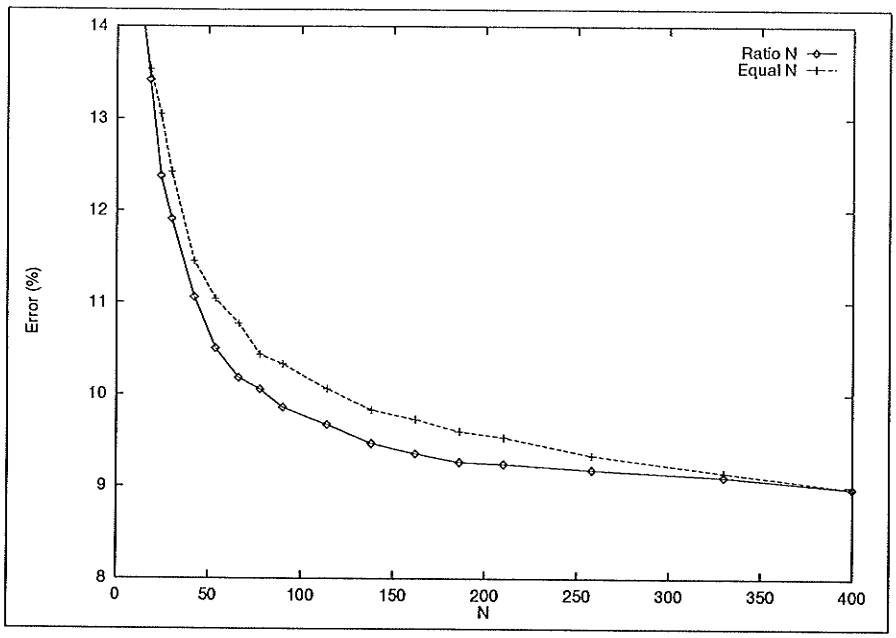

Figure 3.5: Classification results of the Gaussian Data Classification II using the Radial Kernel Classifier with the Ratio N method and the Equal N method

3.3.3 Results and Discussion

The results of this experiment are plotted in Figure 3.5. These together with the results from the previous experiment suggest that the best number of centroids per class for the RKC does not depend on the a priori probabilities. If it does, then the RNM should always give a smaller error rate than the ENM. Since the a priori probabilities can no longer be used as guideline to determine the number of centroids one should use for each class, the next best choice is to use the ENM, that is, using the same number of centroids for every class. Although ENM may not give the optimum classification error rate, it does save time compared to searching through all combinations of $N^{(u)}$ and it provides one with a guideline especially when the a priori probabilities are not known. Unless it is known in advance that some classes require more centroids than other classes, in order to have a better representation, ENM seems to be the most logical choice. In the rest of the experiments in this thesis, the ENM method is used for selecting the number of centroids for each class.

3.4 Summary

This chapter showed that selecting centroids using OCON is better than MCON, thus the use of OCON is recommended. It was established that the number of centroids for each class does not appear to depend on the a priori probabilities. Since the a priori probabilities cannot be used as guideline for determining the number of centroids that should be used for each class, it is recommended that an equal number of centroids for each class be used.

Chapter 4

Location of Centroids

It is mentioned in Chapter 3 that the number of centroids and their locations are two important parameters in RKC. Now that how the number of centroids should be selected is established, this chapter considers the question of what technique should be used to select the location of centroids. The performance of five different centroid selection schemes:

1. Random Centers,
2. K-Means Clustering,
3. Partition Around Medoid (PAM),
4. Learning Vector Quantization (LVQ), and
5. Decision Surface Mapping (DSM)

are studied. Details of the techniques are given in the following section.

4.1 Centroid Selection Schemes

4.1.1 Random Centers

Random Centers was used by Broomhead and Lowe in [8] 1988 when they first proposed RBFN. This technique selects centroids randomly from the training data without replacement. Although in [14] Lowe had shown that Random Centers could give reasonable performance for RBFN, from the experiments in this research, the RKC did not perform well with the Random Centers.

The Random Centers technique used in the experiments here is different from the one used in RBFN literature in two ways. Firstly, in the RBFN literature, random centers were selected using MCON, but in the experiments OCON and ENM are used. This ensures that each class has an equal number of centroids to represent them. Secondly, each selected centroid is treated as a center rather than just a prototype. That is, after the centroids are selected, each training data is assigned to its closest centroids using the Euclidean metric, and the weight, $\alpha_i^{(u)}$, is set to the number of training data from class u grouped into cluster $\boldsymbol{C}_i$.

4.1.2 K-Means Clustering

K-Means clustering was first used by MacQueen in 1967 [34] with Moody and Darken [9] 1989, were the first to use it to select centroids for RBFN. Almost all RBFN literature use K-Means clustering or variants of K-Means for centroid selection. MacQueen's K-Means clustering consists of the following steps:

1. The first k data units in the data set are taken as clusters of one member each. These data are the initial centroids.

2. Each of the remaining $n-k$ data units are then assigned to the cluster with the nearest centroid. After each assignment, the mean of the gaining cluster (the

one which has just received a new data) is computed and the centroid of the gaining cluster is set to this mean value.

3. After all data units are assigned, the existing cluster centroids are taken as fixed seed points and one more pass is made through the data set assigning each data unit to the nearest seed point.

The K-Means clustering used in the experiments in this research was a convergent variant of MacQueen's K-Means technique. The steps used are as follows:

1. From the data set, k initial centroids are selected randomly without replacement as clusters of one member each.

2. Each of the remaining $n^{(u)} - k$ data units is assigned to the cluster with the nearest centroid. After each assignment, the mean value of the gaining cluster is computed. The centroid of the gaining cluster is set to this mean value.

3. Each data unit is then taken in sequence and its distances to all centroids are computed. If the cluster with the nearest centroid is not the same as the parent cluster, then this data unit is reassigned to the cluster with the nearest centroid and the centroids of the losing and the gaining clusters are updated to the mean value of the corresponding cluster.

4. Step 3 is repeated until convergence is achieved. In other words, the clustering process stops when a full cycle through the data set fails to cause any changes in the cluster membership.

Since OCON is used in the experiments, the k in the convergent K-Means algorithm is the number of centroids per class and the $n^{(u)}$ is the number of training data from class u.

The reason for studying K-Means clustering in this thesis is because it is a standard technique that every RBFN paper used. Although using K-Means convergent

clustering with RKC does not outperform other techniques that are studied in the experiments, it gives good performance and its learning speed is faster than those of the LVQ, DSM and PAM.

4.1.3 Partition Around Medoid (PAM)

The PAM technique used in the experiments is described in the book [37] by Kaufman and Rousseeuw. To date no RBFN paper has used this technique to locate the centroids. This technique is very similar to the K-Means clustering. Both try to select centroids which would minimize the sum of distances between the centroids and the data. The difference is that using PAM, the centroids can only be selected from the training data. In other words, the centroids selected have to be one of training data points.

The algorithm of PAM consists of two phases, the BUILD phase and the SWAP phase. In the BUILD phase, an initial clustering is obtained by the successive selection of representative data until k data have been found. It contains the following steps:

1. Select a data which has the smallest sum of distances to all other data. This is the first initial centroid.

2. Consider another data $\boldsymbol{x}_i$ which has not yet been selected.

3. Consider a non selected data, $\boldsymbol{x}_j$. Calculate, D_j, the distance between $\boldsymbol{x}_j$ and its nearest centroid, and $d(i,j)$, the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

4. If the difference between D_j and $d(i,j)$ is positive, then data $\boldsymbol{x}_j$ will contribute

$$C_{ji} = \max(D_j - d(j,i), 0) \tag{4.1}$$

   to the decision of whether data $\boldsymbol{x}_i$ should be selected.

5. Calculate the total gain, $\sum_j C_{ji}$.

6. Select the data $\boldsymbol{x}_i$ which has the maximum total gain as one of the initial centroids.

7. Repeat step 2 to step 6 until k centroids are found.

Next is the SWAP phase. In this second phase, PAM will try to further minimize the sum of distances between the centroids and the data. This is done by considering all pairs of data $(\boldsymbol{x}_\ell, \boldsymbol{x}_m)$ for which data $\boldsymbol{x}_\ell$ is a centroid and data $\boldsymbol{x}_m$ is not and to determine what effect a swap would have on the value of the clustering. The SWAP phase has the following steps:

1. Consider a non selected data $\boldsymbol{x}_k$ and calculate its contribution $C_{k\ell m}$ to the swap:

   (a) If $\boldsymbol{x}_k$ is further away from both $\boldsymbol{x}_\ell$ and $\boldsymbol{x}_m$ than from one of the other centroids, then $C_{k\ell m}$ is zero.

   (b) If $\boldsymbol{x}_\ell$ is the nearest centroid of $\boldsymbol{x}_k$ ($d(k,\ell) = D_k$), then two situations must be considered:

      (i) $\boldsymbol{x}_k$ is closer to $\boldsymbol{x}_m$ than to the second closest centroid

$$d(k,m) < E_k \tag{4.2}$$

      where E_k is the distance between $\boldsymbol{x}_k$ and the second nearest centroid. In this case the contribution of data $\boldsymbol{x}_k$ to the swap between data $\boldsymbol{x}_\ell$ and $\boldsymbol{x}_m$ is

$$C_{k\ell m} = d(k,m) - d(k,\ell). \tag{4.3}$$

      (ii) $\boldsymbol{x}_k$ is at least as distant from $\boldsymbol{x}_m$ as from the second nearest centroid

$$d(k,m) \geq E_k \tag{4.4}$$

      In this case the contribution of object $\boldsymbol{x}_k$ to the swap is

$$C_{k\ell m} = E_k - D_k. \tag{4.5}$$

Note that in situation (i) the contribution $C_{k\ell m}$ can be either positive or negative depending on the relative position of data $\boldsymbol{x}_k$, $\boldsymbol{x}_\ell$, $\boldsymbol{x}_m$. Only when the data $\boldsymbol{x}_k$ is closer to $\boldsymbol{x}_\ell$ than to $\boldsymbol{x}_m$ will the contribution be positive. This indicates that the swap is not favorable from the point of view of data $\boldsymbol{x}_k$. On the other hand, in situation (ii) the contribution is always positive because it cannot be advantageous to replace $\boldsymbol{x}_\ell$ by a data $\boldsymbol{x}_m$ which is further away from $\boldsymbol{x}_k$ than from the second closest centroid.

(c) $\boldsymbol{x}_k$ is further away from data $\boldsymbol{x}_\ell$ than from at least one of the other centroid but closer to $\boldsymbol{x}_m$ than to any centroid. In this case the contribution of $\boldsymbol{x}_k$ to swap is

$$C_{k\ell m} = d(k, m) - D_k. \tag{4.6}$$

2. Calculate the total effect of a swap by adding the contributions $C_{k\ell m}$:

$$T_{\ell m} = \sum_k C_{k\ell m} \tag{4.7}$$

3. Select the pair of $(\boldsymbol{x}_\ell, \boldsymbol{x}_m)$ which $\min_{\ell,m} T_{\ell m}$.

4. If the minimum $T_{\ell m}$ is negative, the swap is carried out and the algorithm returns to step 1 of the SWAP phase. If the minimum $T_{\ell m}$ is positive or 0, then the algorithm stops.

Note that since all potential swaps are considered, the resulting centroids generated using PAM do not depend on the order of the data.

The reason for including PAM in this thesis is because one would like to study whether a centroid selection technique such as PAM which selects a subset of the training data as centroids could perform as well as other techniques such as K-Means which are not constrained to select its centroids from the training data. From the results of the experiments, it is observed that a RKC which uses PAM had a slower convergent rate than when it is used with K-Means, LVQ or DSM.

4.1.4 Learning Vector Quantization (LVQ)

This technique was first proposed by Kohonen in [38] 1988 and was used in [15, 39, 40] to locate centroids for the RBFN. The LVQ technique implemented in the experiments has the following steps:

1. Select k initial centroids from each class of the training data using the convergent variant of the MacQueen's K-Means technique described in Section 4.1.2.

2. Select a data $\boldsymbol{x}_i$ randomly from the training data.

3. Calculate the distance between $\boldsymbol{x}_i$ and all centroids and select the nearest centroid $\boldsymbol{c}_j$.

4. If $\boldsymbol{c}_j$ and $\boldsymbol{x}_i$ belong to the same class, the centroid $\boldsymbol{c}_j$ is moved towards $\boldsymbol{x}_i$ using the equation

$$\boldsymbol{c'}_j = \boldsymbol{c}_j + \beta(t)(\boldsymbol{x}_i - \boldsymbol{c}_j) \tag{4.8}$$

   where $\boldsymbol{c'}_j$ is the new centroid, $\beta(t)$ is a monotonically decreasing linear function which starts at 0.3 and reaches zero in 100,000 steps and t is the number of times the centroids are trained. If $\boldsymbol{c}_j$ and $\boldsymbol{x}_i$ belong to two different classes then $\boldsymbol{c}_j$ is moved away from $\boldsymbol{x}_i$ using the equation

$$\boldsymbol{c'}_j = \boldsymbol{c}_j - \beta(t)(\boldsymbol{x}_i - \boldsymbol{c}_j). \tag{4.9}$$

   Only the nearest centroid is updated.

5. Step 2 to 4 are repeated 100,000 times.

The reason for studying LVQ is because one would like to observe whether LVQ can improve the centroid locations generated by K-Means clustering. In the experiments, it is observed that LVQ outperformed K-Means slightly in the first experiment and it gives the same performance as the K-Means in the other two experiments. Although

LVQ can select better centroids than the K-Means, its learning time was more than double the time used by the K-Means. As a result, the use of LVQ for centroid selection is not recommended.

4.1.5 Decision Surface Mapping (DSM)

DSM was first proposed in the paper [41] by Geva and Sitte in 1991 as a technique to select prototypes for the Nearest Neighbor Classifier. In Geva and Sitte's paper, they showed that DSM is better than LVQ for the pattern classification problem. This technique has not been used for centroid selection in any RBFN paper.

DSM is actually a variation of the LVQ method. The only difference between DSM and LVQ is that DSM does not require the centroids to reflect the probability distribution of each class. Instead, DSM adapts the centroids to closely map the decision surface or boundary which separates the classes. The DSM algorithm consists of the following steps:

1. Select k initial centroids from each class of the training data using the convergent variant of the MacQueen's K-Means technique described in Section 4.1.2.

2. Select a data $\boldsymbol{x}_i$ randomly from the training data.

3. Calculate the distances between $\boldsymbol{x}_i$ and all centroids and select the nearest centroid $\boldsymbol{c}_j$.

4. If $\boldsymbol{c}_j$ and $\boldsymbol{x}_i$ belong to the same class then no modification is made. If they belong to two different classes then $\boldsymbol{c}_j$ is punished and the nearest centroid $\boldsymbol{c}_k$ which belongs to the same class as $\boldsymbol{x}_i$ is rewarded. The centroid $\boldsymbol{c}_j$ is moved away from $\boldsymbol{x}_i$ using the following formula:

$$\boldsymbol{c}'_j = \boldsymbol{c}_j - \beta(t)(\boldsymbol{x}_i - \boldsymbol{c}_j) \tag{4.10}$$

where $\boldsymbol{c}'_j$ is the new centroid, $\beta(t)$ is a monotonically decreasing linear function which starts at 0.3 and reaches zero in 100,000 steps and t is the number of times that the centroids are trained. The centroid $\boldsymbol{c}_k$ is moved toward $\boldsymbol{x}_i$ using the following formula:

$$\boldsymbol{c}'_b = \boldsymbol{c}_b - \beta(t)(\boldsymbol{x}_i - \boldsymbol{c}_b) \tag{4.11}$$

5. Steps 2 to 4 are repeated 100,000 times.

The reason for studying DSM is because one would like to observe the importance of the probability distribution of each class towards the selection of centroids. From the experiments, it is learned that the probability distributions are important to the selection of centroids only when these distributions have simple forms (such as the uniform distribution). As the probability distributions become more and more complex (such as those in the Vowel Experiment), their importance in the selection of centroids decreases.

4.2 Experiments

Three experiments were used to study the performance of the five centroid selection techniques. The first one is the Waveform experiment used in Section 3.2.1. The second experiment is a three-class, two dimensional classification problem with uniformly distributed data. The third experiment is the Vowel Classification experiment described in Section 3.2.2. In these experiments, OCON and ENM are used to select N, the Gaussian Kernel is used as the radial kernel function, and the Euclidean metric is used. Details of these experiments are as follows.

4.2.1 Uniform Data Classification

This first experiment is a three-class, 2 dimensional classification problem with uniformly distributed random data.

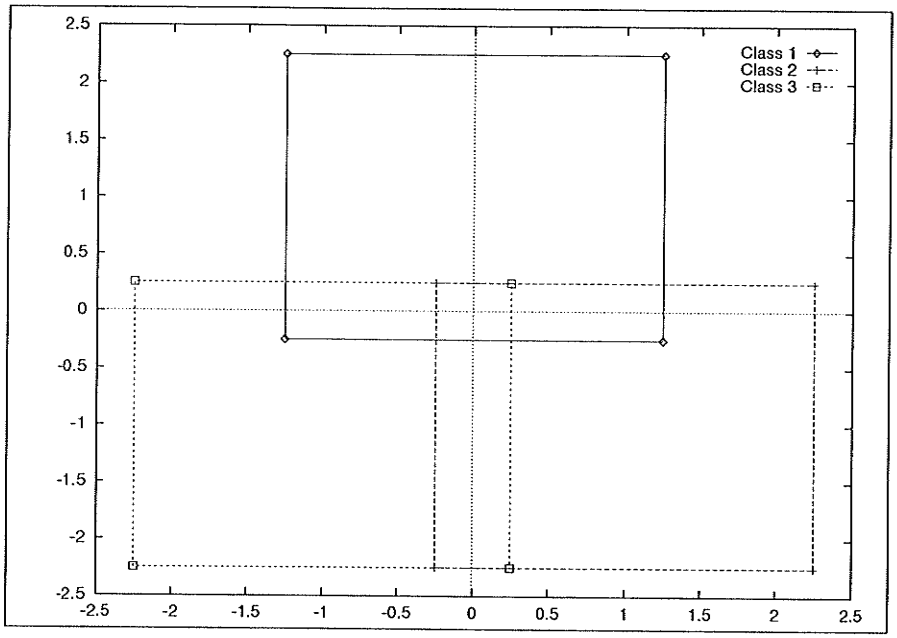

Figure 4.1: Boundary of the Three-Class Uniform Data

Procedures

The three classes of data were generated using the following method:

Class 1 : Generated by randomly selecting a pair of coordinates in the square with corner points $(-1.25, -0.25)$ and $(1.25, 2.25)$.

Class 2 : Generated by randomly selecting a pair of coordinates in the square with corner points $(-0.25, -2.25)$ and $(2.25, 0.25)$.

Class 3 : Generated by randomly selecting a pair of coordinates in the square with corner points $(-2.25, -2.25)$ and $(0.25, 0.25)$.

The boundary of these classes is shown in Figure 4.1.

Three hundred training data per set were generated using the a priori probabilities of $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Thus there were about 100 training data per class. Ten training data sets were generated using the above procedure together with ten test sets of size 3000.

For each set of data, centroids are selected independently from each class using one of the five centroid selection schemes. Then the weight, $\alpha_i^{(u)}$, is set to the number of training data from class u grouped into cluster $\boldsymbol{C}_i$. Lastly, the smoothing parameter h is optimized using the test set. A list of classification errors in percentage and the corresponding h were recorded. Since all centroid selection schemes, except PAM, depend on the location of the initial centroids, these initial centroids were selected randomly from the training data and a RKC was trained ten times for each training and testing set in order to minimize the effect of the dependency. As a result, the classification errors recorded here are an average over ten repetitions for each data set and over the ten sets of training and testing data. The results of this experiment are reported in the next section.

Results

The error rate versus the number of centroids N curves of the five centroid selection schemes are plotted in Figure 4.2. From these results, it is clear that Random Centers cannot compete with other four techniques. Although the error rate of DSM is eventually the same as those of the K-Means, PAM and LVQ, the converging rate of the DSM is slow compared to these three techniques. As a note, the classification error rate that all five techniques converged to was the error rate of the Classical Kernel Classifier.

Discussion

From the results of this uniform data experiment, there are three observations. The first observation is that RKC does not have a good performance if one simply selects a random subset from the training data as centroids (as is the case of Random Centers). The second observation is that although using a random subset of the training data as centroids is undesirable, if one carefully selects a subset of the training data as

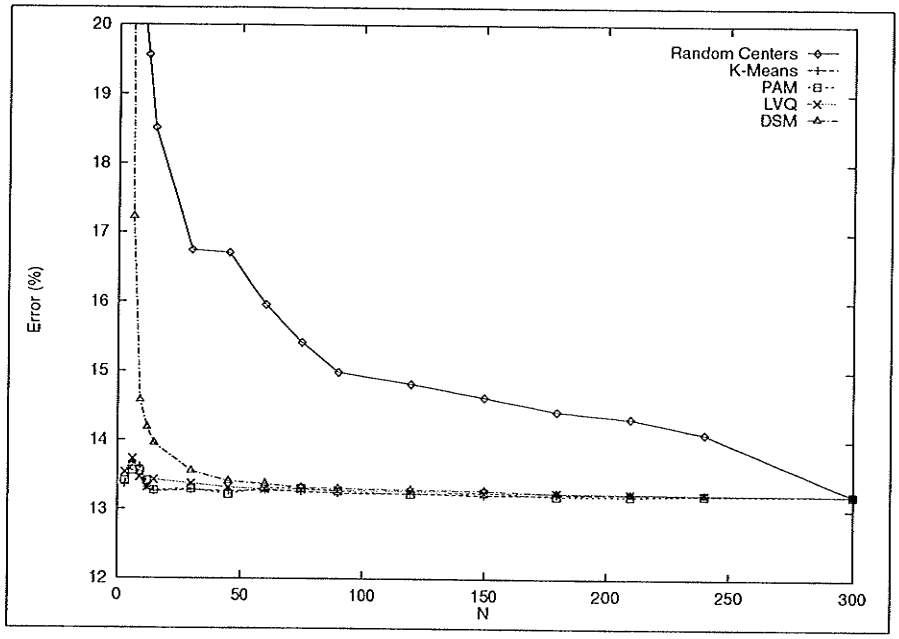

Figure 4.2: Classification results of the Uniform Data Experiment using Radial Kernel Classifier with five centroid selection schemes

centroids using a technique such as the PAM then one could obtain results comparable to those obtained by using K-Means. A third observation is that using centroids which do not reflect the probability distributions of classes may slow down the error convergent rate as in the case of DSM. This suggests that the probability distributions are important to the selection of centroids when N is small. From these observations, the use of K-Means clustering is recommended for centroid selection because of its speed, convergence rate and accuracy.

4.2.2 Waveform Classification

This experiment is the three-class, 21 dimensional waveform classification problem described in Section 3.2.1. For the procedures of this experiment, please refer to that section.

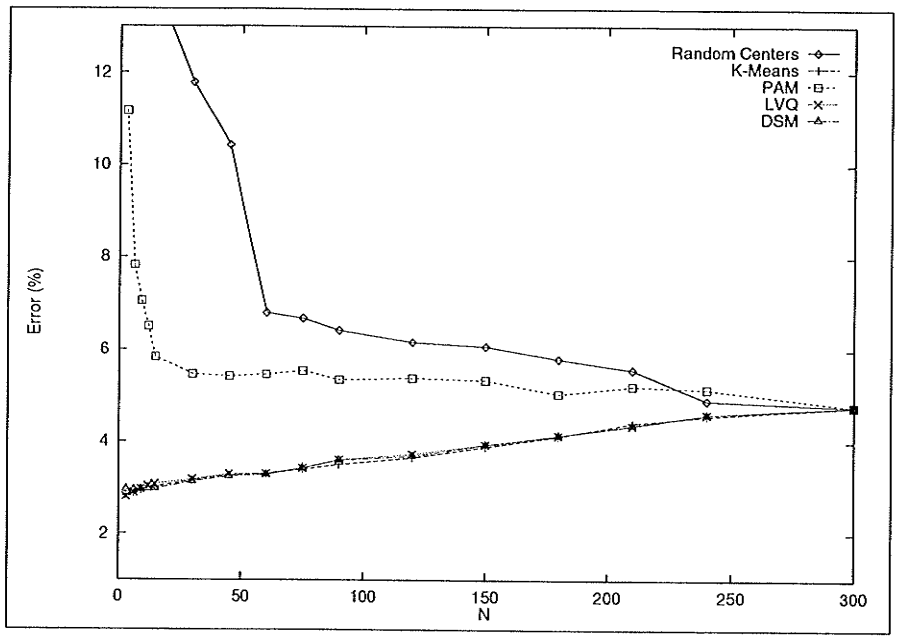

Figure 4.3: Classification results of the Waveform Experiment using Radial Kernel Classifier with five centroid selection schemes

Results

The error rate versus N of the five centroids selection schemes are plotted in Figure 4.3. Similar to the results of the previous experiment, the Random Centers had the worst classification performance, while LVQ and K-Means had similar performance. Unlike the previous results however, PAM was not able to perform as good as DSM, LVQ and K-Means in this experiment.

Discussion

From these results, there are two observations. The first observation is that PAM does not perform as well as DSM, LVQ and K-Means. This is because PAM is constrained to select a subset of the training data as centroids. This result suggests that if one is limited to use a subset of the training data as centroids, then the

error rate will not be better than the error rate of the CKC. The second observation is that in this experiment DSM has a similar performance to that of LVQ and K-Means. Although DSM appears to have difficulties in locating good centroids when the data have a simple distribution (such as Uniform Distribution in the previous experiment), the DSM performs as well as LVQ and K-Means when the data have complicated distributions (such as those in this experiment). This suggests that as the probability distributions of the data class become more complex, their influence on the locations of the centroids and the error rate of the RKC become less.

4.2.3 Vowel Classification

This experiment is the ten-class vowel classification problem described in Section 3.2.2. A list of the vowel data used in this experiment is in Appendix A. For the procedures of this experiment, please refer to Section 3.2.2.

Results

The error rate versus the number of centroids N curves of the RKC using the five centroid selection techniques are plotted in Figure 4.4. The most interesting result of this experiment is that DSM outperforms LVQ, PAM and K-Means. The remaining results are similar to those of the previous two experiments.

Discussion

The most surprising result in this real data experiment was that DSM had outperformed PAM, K-Means, LVQ and even Classical Kernel Classifier. It appears that as the probability distributions of classes become complicated, techniques such as K-Means and LVQ start to have trouble selecting centroids which model these distributions. By having the freedom to select centroids which do not necessary reflect the distribution of the data, DSM is able to focus its attention on selecting centroids

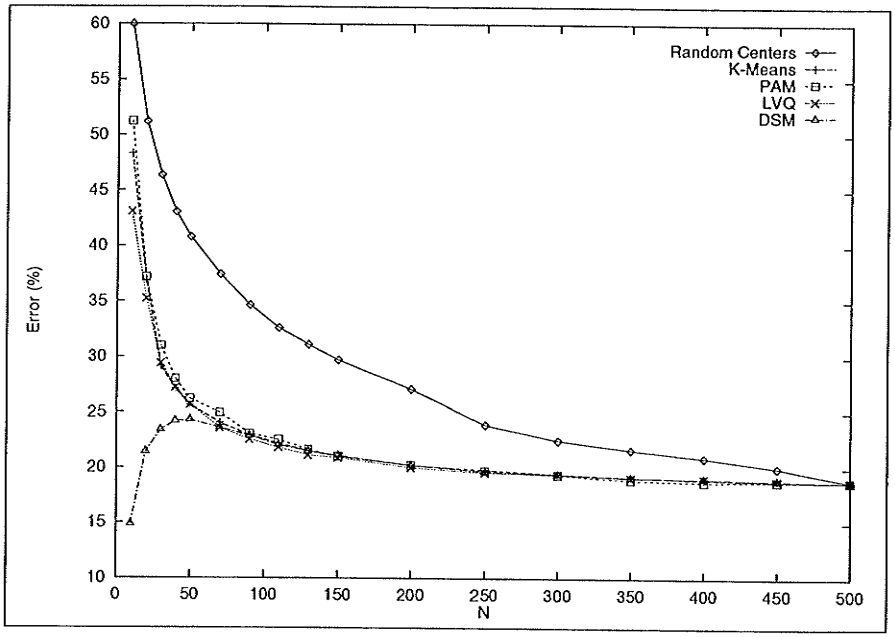

Figure 4.4: Classification results of the Vowel Experiment using Radial Kernel Classifier with five centroid selection schemes

which improve classification rather than on modeling the distribution. Thus, it was able to achieve the best error rate in this experiment.

4.3 Summary

From the results presented in this chapter, there are four important observations.

1. Using Random Centers as RKC's centroids usually results in an error rate that is worse than those obtained by the other four techniques. Thus using the Random Centers with the RKC is not recommended.

2. Although the results of the PAM in the first and the third experiment are similar to those obtained by using the K-Means and the LVQ, the constraint which limits the PAM to select centroids from only the training data slows down

PAM's convergence rate as in the case of the Waveform Experiment. Unless the problem requires the use of a subset of training data as centroids, PAM is not recommended.

3. Since the K-Means gives similar results to the LVQ and since it has a shorter learning time than the LVQ in all three experiments, the use of the K-Means clustering for centroid selection in RKC is recommended.

4. Forcing the centroids to model the probability distribution of the classes may not necessarily provide good classification performance. A good example is the surprisingly good results obtained by DSM – a technique which does not constrain its centroids to reflect the probability distribution of the classes – in the Vowel Experiment. By outperforming the other four techniques in this last experiment, DSM shows its potential for centroid selection in real applications. However, the use of the DSM for centroid selection is not recommended because its learning time is longer than those of the K-Means and it has a slow convergence rate.

Based on the third observation, the K-Means clustering is used for centroid selection in the remaining experiments of this research.

Chapter 5

Distance Measures

This chapter studies the performance of RKC when six different metrics are used. In order to simplify the study, the six metrics are grouped into two sets: the L_p metric and the Mahalanobis metric. The L_p metric consists of the L_1 metric or the Manhattan distance, the Euclidean metric and the L_∞ metric. The Mahalanobis metric consists of the Global Sigma metric, the One Class One Sigma (OCOS) metric and the P-Nearest Neighbor metric. For each metric group, two experiments were used in to study their performance. The Waveform and the Vowel experiments were used to study the performance of RKC using an L_p metric, and the Uniform Data and the Vowel experiments were used to study the performance of an Mahalanobis metric.

5.1 L_p Metric

In this section, the performance of the RKC with three different L_p metrics is compared.

1. The Manhattan metric or L_1 norm which has the form:

$$\|\boldsymbol{x}-\boldsymbol{y}\| = \sum_{i=1}^{d} |x_i - y_i|, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{R}^d. \tag{5.1}$$

2. The Euclidean metric or L_2 norm which has the form:

$$\|\boldsymbol{x}-\boldsymbol{y}\| = \left\{\sum_{i=1}^{d}(x_i-y_i)^2\right\}^{1/2}, \qquad \boldsymbol{x},\boldsymbol{y}\in\mathcal{R}^d. \tag{5.2}$$

3. The Maximum Value metric or L_∞ norm which has the form:

$$\|\boldsymbol{x}-\boldsymbol{y}\| = \max\{|x_i-y_i|\}_{i=1}^{d}, \qquad \boldsymbol{x},\boldsymbol{y}\in\mathcal{R}^d. \tag{5.3}$$

Both the Manhattan metric and the Maximum Value metric have not been studied in RBFN literature.

The two experiments used to study the performance of RKC are the Waveform experiment described in Section 3.2.1 and the Vowel experiment which is described in Section 3.2.2. In these experiments, OCON and ENM were used to select the number of centroids, K-Means clustering was used for selecting the centroids and the Gaussian Kernel was used as the radial kernel function.

5.1.1 Waveform Classification

This experiment is the three-class, 21 dimensional waveform classification problem described in Section 3.2.1. For the procedures of this experiment, please refer to Section 3.2.1.

Results

The error rate versus N curves for the three metrics are plotted in Figure 5.1. From the results, RKC performs slightly better with the Euclidean metric than with the Manhattan metric. The Maximum Value metric has the worst performance for any N.

Discussion

From the results, there are two observations.

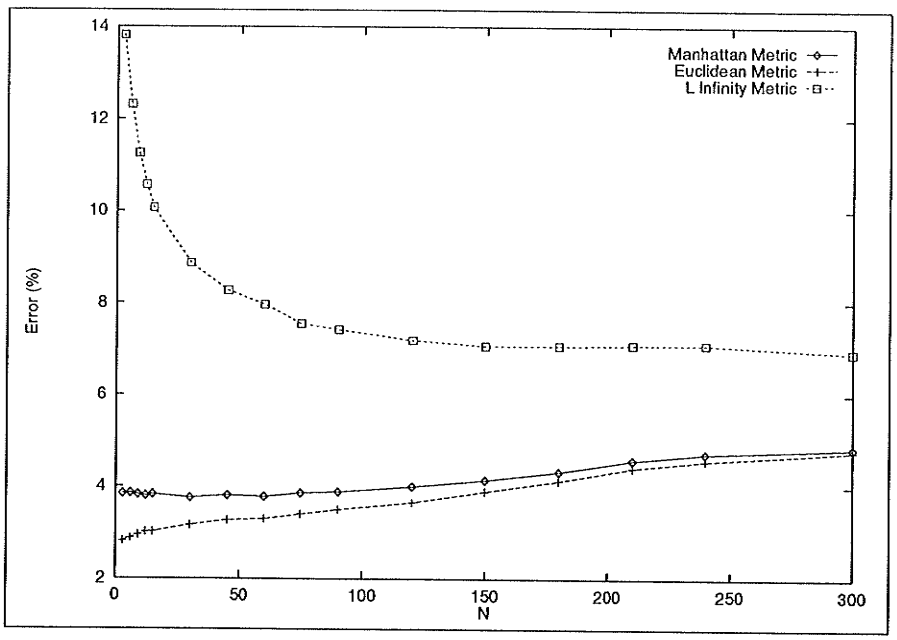

Figure 5.1: Classification results of the Waveform Experiment using Radial Kernel Classifier with three L_p metrics

First, although the Manhattan metric does not outperform the Euclidean metric, the results are very close; within one percent. Since the classification speed of a RKC using the Manhattan metric is faster than that of the Euclidean metric, if classification speed is more important than accuracy, then the Manhattan metric should be used instead of the Euclidean metric.

Second, the Maximum Value metric has the largest error rate for every N. This poor performance of the Maximum Value metric is reasonable because it used only one of the 21 dimensions for its distance measure. In other words, the information within the remaining 20 dimensions was discarded by the Maximum value metric. No wonder it performs poorly.

5.1.2 Vowel Classification

This experiment is the ten-class vowel classification problem described in Section 3.2.2. For the procedures of this experiment, please refer to Section 3.2.2.

Results

The error rate versus N curves for the three different metrics are plotted in Figure 5.2. The most remarkable thing about these results is that a RKC using the Manhattan metric has actually performed better than a RKC using the Euclidean metric.

Discussion

This experiment shows that for RKC the Manhattan metric is a viable alternative to the use of the Euclidean metric. The Manhattan metric is both faster than the Euclidean metric to compute and it provides a better performance than the Euclidean metric.

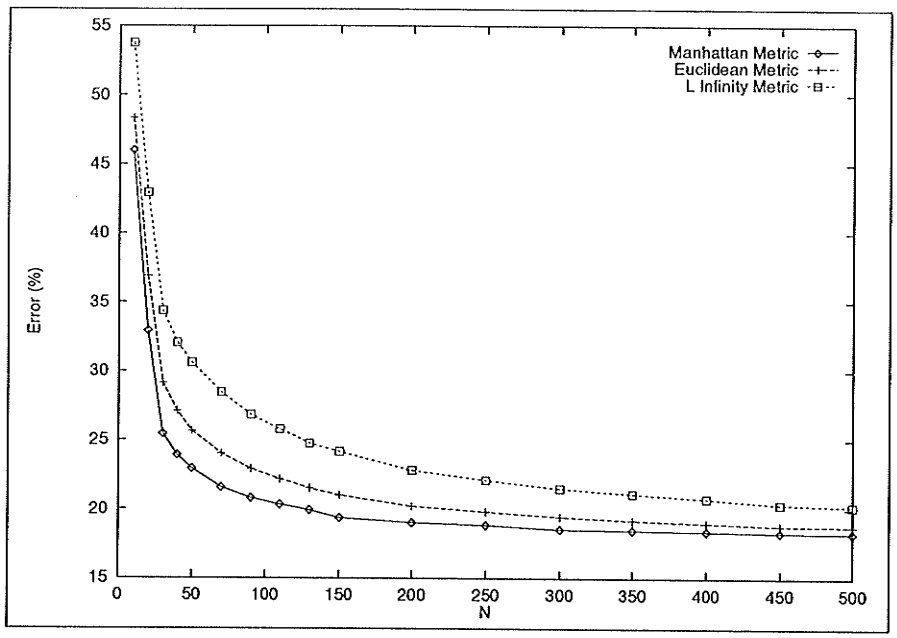

Figure 5.2: Classification results of the Vowel Experiment using Radial Kernel Classifier with three L_p metrics

5.1.3 Summary

These two experiments show that

1. The Maximum Value metric or the L_∞ norm with RKC gives results which are worse than those of the Euclidean metric. Thus its use is not recommended.

2. The use of the Manhattan metric is an attractive alternative to the use of the Euclidean metric. Not only can it provide a classification speed faster than that of the Euclidean metric, but the Manhattan metric can also provide a classification error similar to or even better than the Euclidean metric.

In the next section, the metrics that belong to the Mahalanobis metric group are studied.

5.2 Mahalanobis Metric

The performance of RKC using the Manhattan metric is compared with three different Mahalanobis metrics:

1. the P-Nearest Neighbor metric;

2. the Global Sigma metric; and

3. the One Class One Sigma (OCOS) metric.

Recall that the Radial Kernel Density Estimate (RKDE) is used by the RKC to estimate the class-conditional probability density $p(\boldsymbol{x}|\omega_u)$. It has the form

$$q^{(u)}(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{N^{(u)}} \alpha_i^{(u)} \, \phi\left(\frac{\|\boldsymbol{x} - \boldsymbol{c}_i^{(u)}\|}{h}\right) \tag{5.4}$$

where n is the total number of training data, h is the smoothing parameter, d is the dimension of $\boldsymbol{x}$, $N^{(u)}$ is the number of clusters which contain data from class u, $\alpha_i^{(u)}$

is the number of training data from class u grouped into cluster $\boldsymbol{C}_i$, ϕ is the radial kernel function, and $\boldsymbol{c}_i{}^{(u)}$ is the i-th centroid. When the Mahalanobis metric is used with the RKC, the RKDE will have the form

$$q^{(u)}(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{N^{(u)}} \frac{\alpha_i^{(u)}}{\left(\det \Sigma_i^{(u)}\right)^{1/2}} \phi \left(\frac{\left((\boldsymbol{x} - \boldsymbol{c}_i{}^{(u)})^T (\Sigma_i^{(u)})^{-1} (\boldsymbol{x} - \boldsymbol{c}_i{}^{(u)}) \right)^{1/2}}{h} \right) \tag{5.5}$$

where $\Sigma_i^{(u)}$ is the normalization matrix for the cluster $\boldsymbol{C}_i$, $\det \Sigma_i^{(u)}$ is the determinant of $\Sigma_i^{(u)}$, $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. For each different Mahalanobis metric, the calculation of $\Sigma_i^{(u)}$ is different.

The Uniform Data experiment and the Vowel experiment were used in this section. In these experiments, OCON and ENM were used to select N, the K-Means clustering was used for centroid selection, and the Gaussian Kernel was used as the radial kernel function.

5.2.1 P-Nearest Neighbor Metric

For the P-Nearest Neighbor metric, the normalization matrix $\Sigma_i^{(u)}$ is set to the root mean square value of the Euclidean distances from the centroid $\boldsymbol{c}_i^{(u)}$ to its P nearest neighboring centroids of the same class, that is,

$$\Sigma_i^{(u)} = \left[\frac{1}{P} \sum_{j=1}^{P} \| \boldsymbol{c}_i{}^{(u)} - \boldsymbol{c}_{(j)}^{(u)} \|^2 \right]^{1/2} \boldsymbol{I} \tag{5.6}$$

where $\| \ldots \|$ is the Euclidean distance, $\boldsymbol{I}$ is an identity matrix, and $\left\{ \boldsymbol{c}_{(j)}^{(u)} \right\}_{j=1}^{P}$ are the P nearest neighboring centroids of $\boldsymbol{c}_i{}^{(u)}$. The P-Nearest Neighbor metric was first proposed by Moody and Darken in [16] 1988 to calculate the receptive field width of the RBFN. A unique feature of the P-Nearest Neighbor is that it provides the RKC with an additional parameter P to improve its classification performance. The experiments show that in general $P = 3$ is sufficient to provide a good classification performance. As a result, the Three Nearest Neighbor metric is used in both experiments in this section. Since the Three Nearest Neighbor metric requires every

centroid to have at least three neighboring centroids from the same class, the number of centroids for each class is started at four for both experiments.

5.2.2 Global Sigma Metric

For the Global Sigma metric, the normalization matrix $\Sigma_i^{(u)}$ is set to the sample covariance matrix of all the training data, that is, $\Sigma_i^{(u)} = \Sigma$. Using Global Sigma metric, the RKDE in equation (5.5) can be written as

$$q^{(u)}(\boldsymbol{x}) = \frac{(\det \Sigma)^{-1/2}}{nh^d} \sum_{i=1}^{N^{(u)}} \alpha_i^{(u)} \, \phi \left(\frac{\left((\boldsymbol{x} - \boldsymbol{c}_i{}^{(u)})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{c}_i{}^{(u)}) \right)^{1/2}}{h} \right). \tag{5.7}$$

This metric system is similar to the one proposed by Poggio and Girosi in [42] 1990.

5.2.3 One Class One Sigma (OCOS) Metric

For the One Class One Sigma (OCOS) metric, the normalization matrix is set to the sample covariance matrix of each class, that is, $\Sigma_i^{(u)} = \Sigma^{(u)}$. Using OCOS, equation (5.5) can be written

$$q^{(u)}(\boldsymbol{x}) = \frac{\left(\det \Sigma^{(u)}\right)^{-1/2}}{nh^d} \sum_{i=1}^{N^{(u)}} \alpha_i^{(u)} \phi \left(\frac{\left((\boldsymbol{x} - \boldsymbol{c}_i{}^{(u)})^T (\Sigma^{(u)})^{-1} (\boldsymbol{x} - \boldsymbol{c}_i{}^{(u)}) \right)^{1/2}}{h} \right). \tag{5.8}$$

The purpose of using OCOS is to reduce the contribution from the data dimensions with a large range or value to the distance measure. This idea of normalizing the contribution of each dimension of the data during classification is similar to the idea of normalizing the training data before centroid selection proposed by Lay and Hwang in [19] 1993.

5.2.4 Uniform Data Classification

The first experiment is the three-class uniform data classification experiment described in Section 4.2.1. For the procedures of this experiment, please refer to Sec-

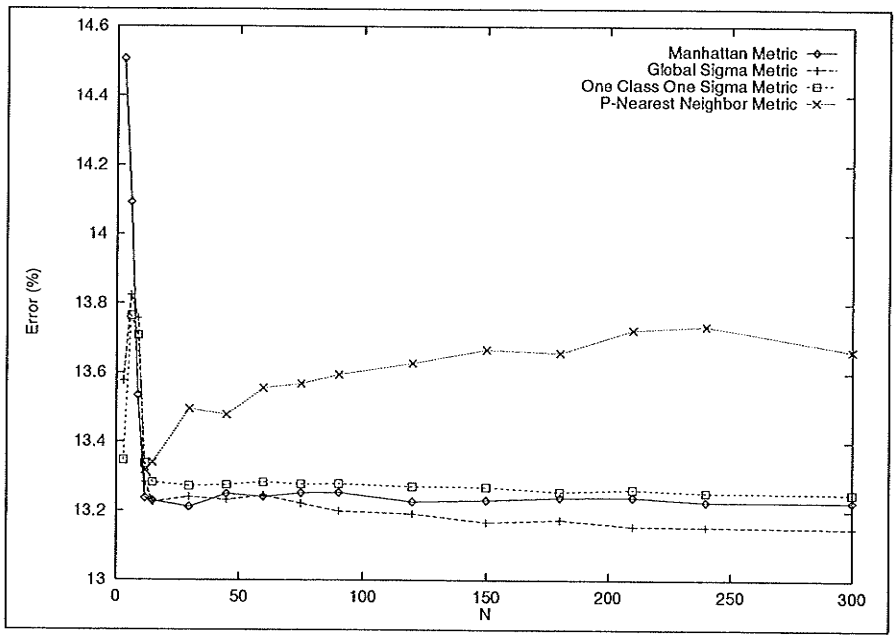

Figure 5.3: Classification results of the Uniform Data Experiment using Radial Kernel Classifier with four Mahalanobis metrics

tion 4.2.1.

Results and Discussion

The error rate versus N curves for the RKC using four metrics are plotted in Figure 5.3. From the results, there are two observations.

The first observation is that although the Global Sigma metric and the OCOS metric are more flexible than the Manhattan metric since they have more parameters to vary, they give similar results to the Manhattan metric in this experiment. The reason for this result is because uniform random data was used. Since RKC using the Global Sigma metric or the OCOS metric have a longer learning and classification time than the Manhattan metric, the use of the Manhattan metric is recommended when the data have an uniform distribution.

The second observation is that the P-Nearest Neighbor metric gives a poor per-

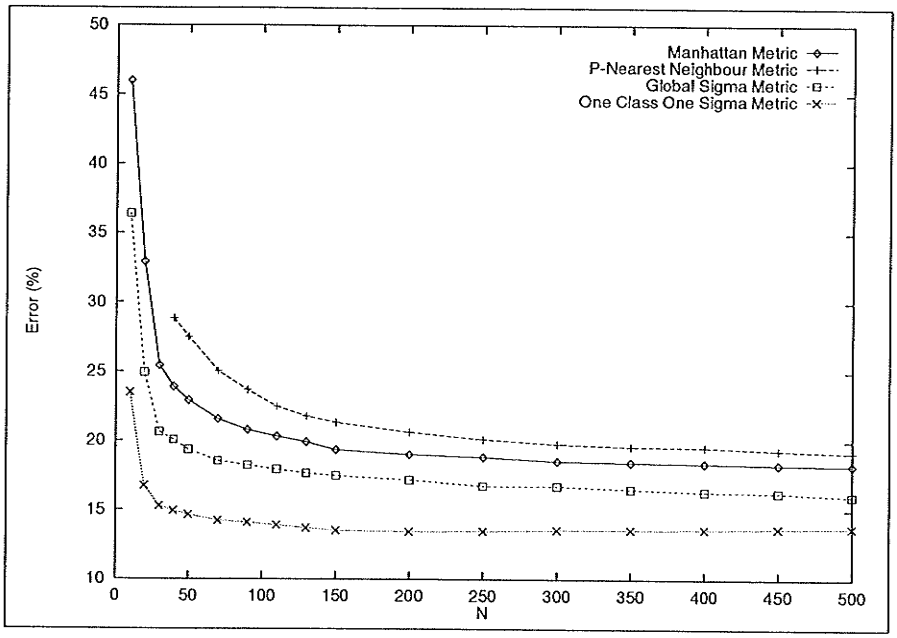

Figure 5.4: Classification results of the Vowel Experiment using Radial Kernel Classifier with four Mahalanobis metrics

formance compared to the other three metrics. Since the P-Nearest Neighbor metric used more local information than other metrics, this result suggests that RKC does not perform well with a local metric.

5.2.5 Vowel Classification

This experiment is the ten-class vowel classification problem described in Section 3.2.2. For the procedures of this experiment, please refer to Section 3.2.2.

Results and Discussion

The error rate versus N curves of the RKC using the four different distance metrics are plotted in Figure 5.4. This experiment illustrates the real advantage of using OCOS and Global Sigma metrics when the data have a wide range and spread in

different dimensions. The reason for their good performance is because the OCOS metric and the Global Sigma metric are able to obtain an objective distance measure by normalizing the Vowel data when the distance measure is calculated.

The results also shows that the P-Nearest Neighbor metric does not improve the classification performance of a RKC using the Manhattan metric. It appears that a local metric such as the P-Nearest Neighbor cannot outperform the Manhattan metric.

5.2.6 Summary

To conclude, the two experiments show that

1. When the data in the classification problem has a large range or spread as in case of the Vowel data, the OCOS metric should be used.

2. Local metrics such as the P-Nearest Neighbor metric should not be used with the RKC because in general they cannot achieve a classification performance better than that of the Manhattan metric.

5.3 Discussion and Summary

If the results of the Vowel Classification experiment in Figure 5.2 are compared with the results in Figure 5.4, it is clear that OCOS outperforms the Manhattan metric. This suggests that OCOS is the best metric to use when the data has a complex distribution (as in the last experiment). The only price one pays for this good performance is a slightly longer classification time. If classification speed is more important than the performance then the Manhattan metric is recommended. Although its performance is not as good as the OCOS's, its classification error is comparable to that of the Euclidean distance metric.

Chapter 6

Radial Kernel Functions

This chapter studies the performance of RKC using 12 different Radial Kernel Functions (RKFs). Since RKC is a hybrid between the Classical Kernel Classifier (CKC) and Radial Basis Functions Network (RBFN), the RKFs considered in this chapter come from the literature of these two classifiers. Note that in the CKC literature the RKF is called Kernel Function and in the RBFN literature it is called Radial Basis Functions. To simplify the study, the 12 RKFs are grouped into three sections:

1. **Second Order Kernel Functions** :

   (a) Gaussian Kernel,

   (b) Rectangular Kernel,

   (c) Epanechnikov Kernel,

   (d) Biweight Kernel, and

   (e) Triangular Kernel.

   They are used in the CKC literature for both pattern classification and function estimation.

2. **Higher-Order Kernel Functions** :

(a) the Fourth Order Kernel,

(b) the Sixth Order Kernel, and

(c) the Eighth Order Kernel.

They are used mainly in the CKC literature for function estimation.

3. **Radial Basis Functions** :

(a) Pseudo-Cubic Spline,

(b) Thin Plate Spline,

(c) Multi-Quadric Equation, and

(d) Logarithmic Basis Function.

They are used in the RBFN literature for function approximation.

The reason for studying the performance of these RKFs is to find out which types of RKFs should not be used with RKC, which may be used, and which should always be used with RKC. The results of the experiments in this chapter show that most RBFs should not be used with RKC, all Kernel Functions may be used with RKC, and the Gaussian Kernel should always be used with the RKC.

Two experiments were used to study the performance of the 12 RKFs. For Second Order Kernels, the Uniform Data and the Vowel experiments were used. Higher-Order Kernels and the Radial Basis Functions used only the Vowel experiment.

6.1 Second Order Kernel Functions

In this section, the performance of the RKC using five second order kernel functions is compared:

1. **Gaussian Kernel** : This is the most widely used kernel in both CKC and RBFN literature and it has the form:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2}. \tag{6.1}$$

2. **Rectangular Kernel** : This is basically the Uniform Distribution function which has the form

$$\phi(x) = \begin{cases} \frac{1}{2} & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

3. **Epanechnikov Kernel** : This kernel has the form:

$$\phi(x) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5}x^2\right) & \text{for } |x| < \sqrt{5}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

4. **Biweight Kernel** : This kernel has the form:

$$\phi(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2 & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

5. **Triangular Kernel** : This kernel has the form:

$$\phi(x) = \begin{cases} 1 - |x| & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

A plot of these kernels is in Figure 6.1. These five kernel functions are similar in that they are all probability density functions. That is

$$\int_{x=-\infty}^{\infty} \phi(x) = 1 \qquad \text{and} \qquad \phi(x) \geq 0 \qquad \forall x$$

for these five kernels. Except for the Gaussian Kernel, each of these second order kernel functions have a finite support, that is, $\phi(x) > 0$ for only a finite range. The purpose of this section is to find out whether a finite support kernel could perform as well as the Gaussian Kernel which has infinite support. Despite the wide acceptance

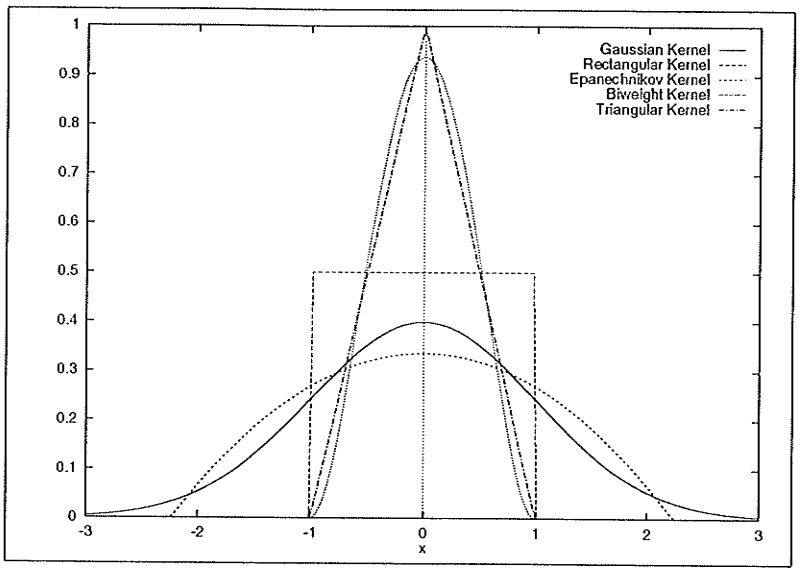

Figure 6.1: Second Order Kernel Functions

of the Gaussian Kernel, its classification speed is slow when compared to the other four Second Order Kernels. The reason for this slow classification speed is because the exponential function in the Gaussian Kernel requires a long computational time and because of the infinite support the Gaussian Kernel has to use this function for every real x. If one of the finite support kernels could produce similar results to the Gaussian Kernel, using it with the RKC classification would mean a significant increase in classification speed.

Two experiments were used to study the performance of the RKC using the five Second Order Kernels. They were the Uniform Data experiment described in Section 4.2.1 and the Vowel experiment described in Section 3.2.2.

6.1.1 Uniform Data Classification

The first experiment is the three-class Uniform Data classification experiment described in Section 4.2.1. For the procedures of this experiment, please refer to Section 4.2.1.

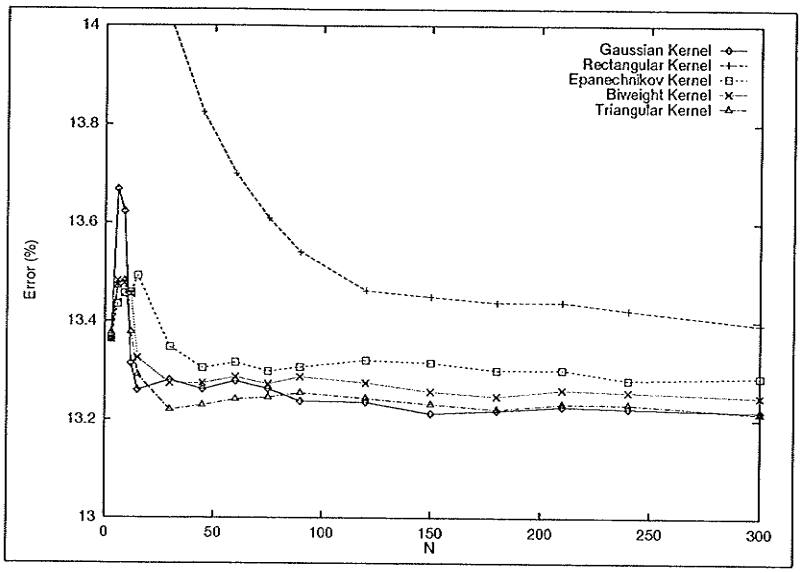

Figure 6.2: Classification results of the Uniform Data Experiment using Radial Kernel Classifier with five Second Order Kernel Functions

Results and Discussion

The results are plotted in Figure 6.2. From the results, there are two observations.

The first observation is that RKC gives a poor performance when it is used with the Rectangular Kernel which is a non-local kernel. A non-local kernel is a kernel function which does not have a convex shape. It appears that the non-local kernel has a lesser discriminating power than the local ones. This fact will become clear if one considers the following example. Consider a two-class classification problem where each class has the same a priori probability and each class has a triangular distribution (Figure 6.3). Using the Bayes Decision Rule (described in Section 2.1), one can achieve the Bayes Error if any observation less than 0 is assigned to class 1 and any observation larger than or equal 0 is assigned to class 2. Now if one tries to estimate the distribution of each class with a Rectangular Kernel (Figure 6.4), the observations which fall between the range $[-0.5, 0.5]$ cannot be properly classified because the two Rectangular Kernels give the same response in this region. This is

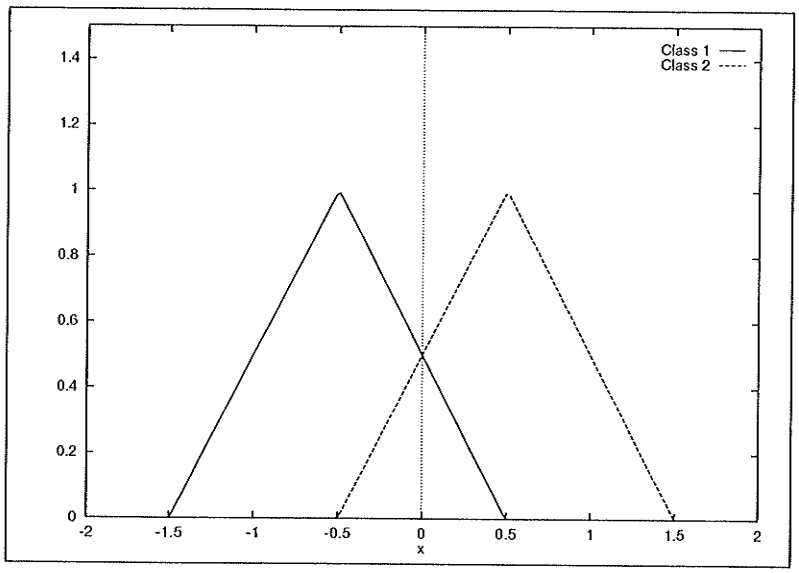

Figure 6.3: Two-Class Problem with Triangular Distribution

exactly what happens when the Rectangular Kernel is used with a RKC. The region of unknown classification in the boundaries between classes is the reason for the poor performance of the RKC using the Rectangular Kernel.

The second observation is that using the Triangular Kernel is an attractive alternative to the use of the Gaussian Kernel. Not only is the Triangular Kernel faster and easier to calculate than the Gaussian Kernel, the Triangular Kernel also achieved a better result than the Gaussian Kernel in this experiment (Figure 6.2). The use of a finite support kernel such as the Triangular Kernel is recommended if the speed of classification is important.

6.1.2 Vowel Classification

This experiment is the ten-class vowel classification problem described in Section 3.2.2. For the procedures of this experiment, please refer to that section.

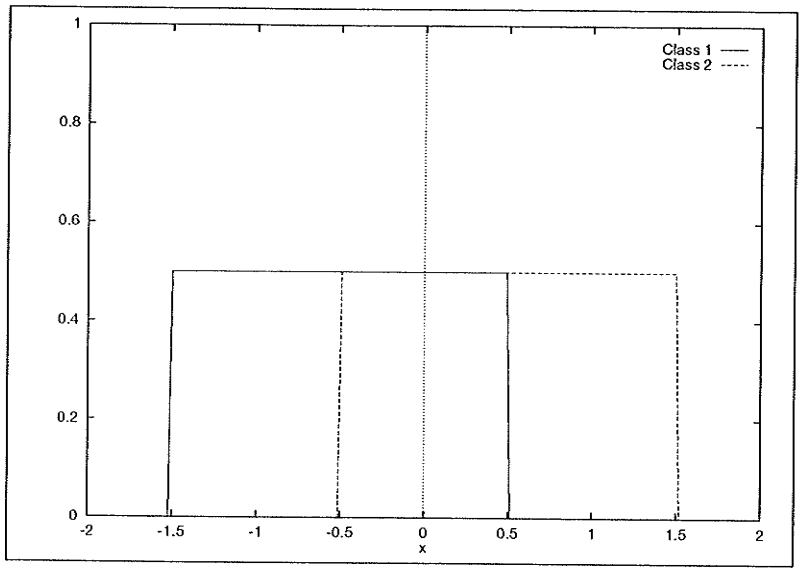

Figure 6.4: Estimate Triangular Distributions with Rectangular Kernels

Results and Discussion

The results are plotted in Figure 6.5. The results of this experiment prompt three observations.

The first observation is that the Rectangular Kernel has the worst performance of the five second order kernels that are studied. This result suggests that the Rectangular Kernel has a lesser discriminating power than the other four kernels. As mentioned in the previous experiment, the reason for this poor performance is because it is hard to get clear boundaries between classes when the Rectangular Kernel is used with a RKC.

The second observation is that the Gaussian Kernel which has an infinite support performs better than the remaining four finite support kernels. The reason for this is probably because RKC which uses an infinite support kernel such as the Gaussian Kernel is able to classify any data in the sample space but when RKC uses a finite support kernel, it cannot give any response to those data that are outside the support of the kernel. As a result, using a finite support kernel with RKC, there may be

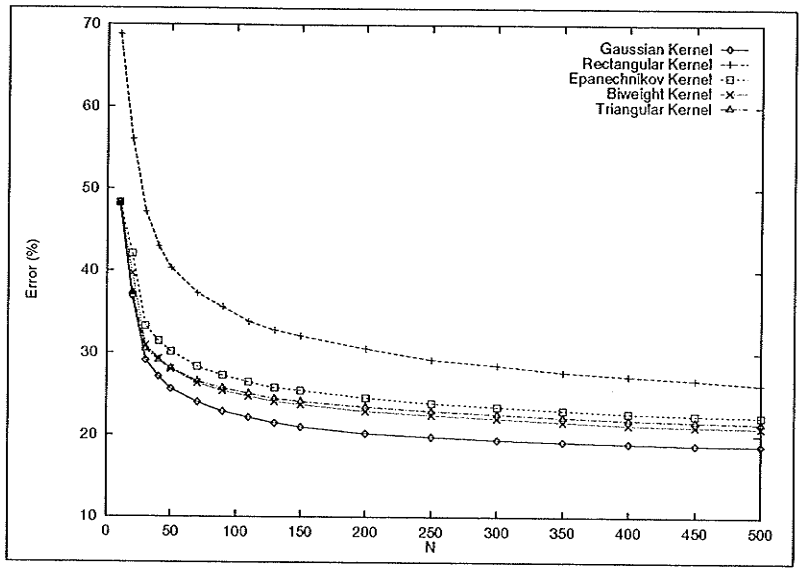

Figure 6.5: Classification results of the Vowel Experiment using Radial Kernel Classifier with five Second Order Kernel Functions

regions in the data space where no radial kernel can provide any response. Data within this region of void therefore cannot be classified. This may explain why the Gaussian Kernel is the most widely used kernel in both the CKC and the RBFN literature.

The third observation is that the three finite support local kernels: the Triangular Kernel, the Biweight Kernel and the Epanechnikov Kernel, gives similar results (Figure 6.5) even though they have a different profile. This suggests that when one wants to use a local kernel for its fast classification speed, one should choose the Triangular Kernel because it has the simplest form and it is the easiest to compute.

6.1.3 Summary

These two experiments show that

1. The Rectangular Kernel should not be used with RKC in classification because its performance is usually worse than those of the other four second order kernels.

2. The Gaussian Kernel should be used with RKC for most classification problems because its results are as good as or even better than those of the other four kernels.

3. If classification speed is important, one should consider using a finite support local kernel at the price of losing some classification performance. Since the three finite support local kernels gave similar results in the experiments and Triangular Kernel has the simplest form in the four local kernels, using the Triangular Kernel is recommended to increase the classification speed of a RKC.

6.2 Higher-Order Kernel Functions

In this section, the performance of a RKC using the Gaussian Kernel is compared with three higher-order kernel functions:

1. **Fourth Order Kernel** : This kernel has the form:

$$\phi(x) = \begin{cases} \frac{15}{32}(1-x^2)(3-7x^2) & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

2. **Sixth Order Kernel** : This kernel has the form:

$$\phi(x) = \begin{cases} \frac{105}{256}(1-x^2)(5-30x^2+33x^4) & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

3. **Eighth Order Kernel** : This kernel has the form:

$$\phi(x) = \begin{cases} \frac{315}{4096}(1-x^2)(35-385x^2+1001x^4-715x^6) & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

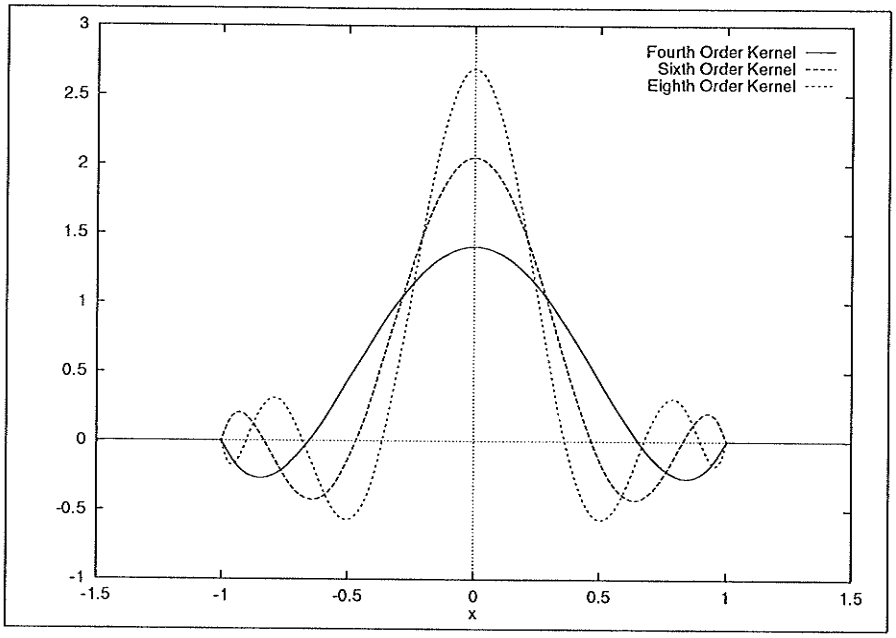

Figure 6.6: Higher-Order Kernel Functions

A plot of the three higher-order kernels is given in Figure 6.6. The main difference between these three higher-order kernels and the second order kernels is that these higher-order kernels have negative responses, that is, $\phi(x)$ could be negative. If the higher-order kernels is used with the Kernel Density Estimate to estimate the class-conditional probability distribution, the resulting distribution will have negative values. This is undesirable because a probability distribution should always be positive. Due to this reason, higher-order kernels are not used in CKC literature. These kernels however are used in Kernel Smoothing or function estimation because the higher the kernel order the faster the convergence rate. The purpose of this section is to find out whether higher-order kernels can outperform the second order kernel such as the Gaussian Kernel.

The Vowel experiment described in Section 3.2.2 is used to compare the performance of the RKC using these three Higher-Order Kernels with the Gaussian Kernel.

The results are plotted in Figure 6.7. They show that the higher the kernel order

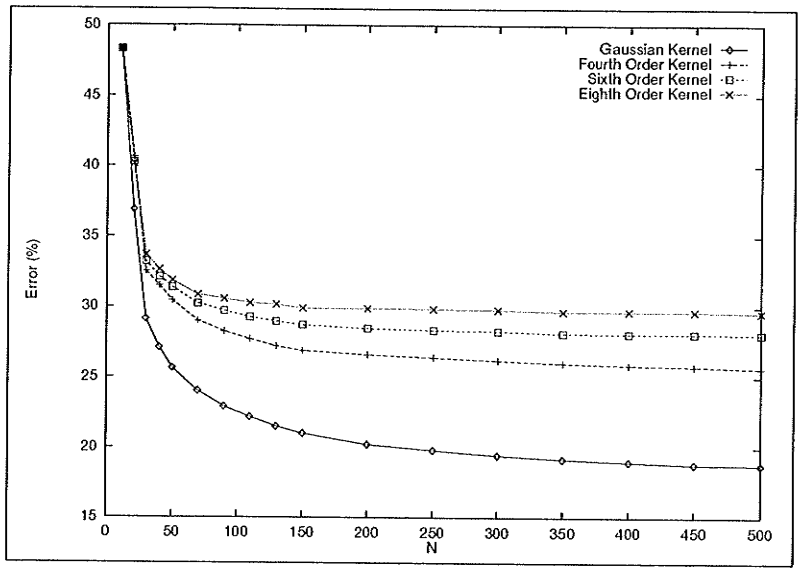

Figure 6.7: Classification results of the Vowel Experiment using Radial Kernel Classifier with Higher Order Kernel Functions

is the poorer is the performance. Thus the Higher-Order Kernels should not be used with the RKC in classification problems.

6.3 Radial Basis Functions

This section studies the performance of a RKC using four radial basis functions:

1. **Pseudo-Cubic Spline (PCS)** : This basis function has the form:

$$\phi(x) = |x|^3. \tag{6.9}$$

2. **Thin Plate Spline (TPS)** : This basis function has the form:

$$\phi(x) = x^2 log(x). \tag{6.10}$$

3. **Multi-Quadric Equation (MQE)** : This basis function has the form:

$$\phi(x) = \sqrt{x^2 + k^2}. \tag{6.11}$$

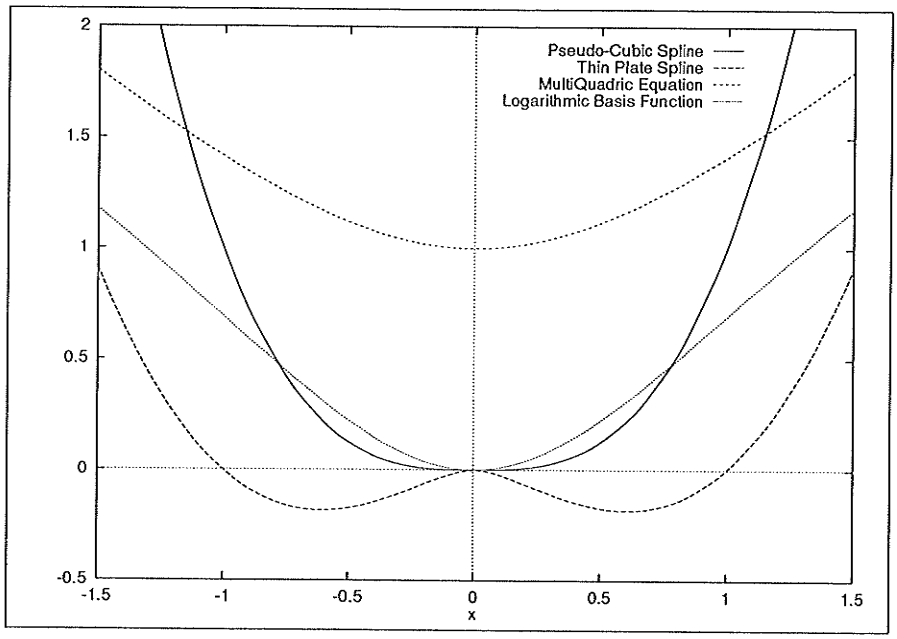

Figure 6.8: Radial Basis Functions

In the experiments, $k = 1$ is used.

4. **Logarithmic Basis Function (LBF)** : This basis function has the form:

$$\phi(x) = \log(x^2 + k^2). \tag{6.12}$$

In the experiments, $k = 1$ is used.

A plot of these radial basis functions is in Figure 6.8. The major difference between these Radial Basis Functions (RBFs) and the kernel functions that were studied in the previous sections is that these have a concave profile.

Again the Vowel experiment described in Section 3.2.2 is used here to study the performance of the RKC using the four Radial Basis Functions.

The results are plotted in Figure 6.9. From these results, it is clear that RKC loses all its classification power when used with the concave RBFs. The reasons for this poor performance lie in the the shape of the RBFs and the weights of the RKC.

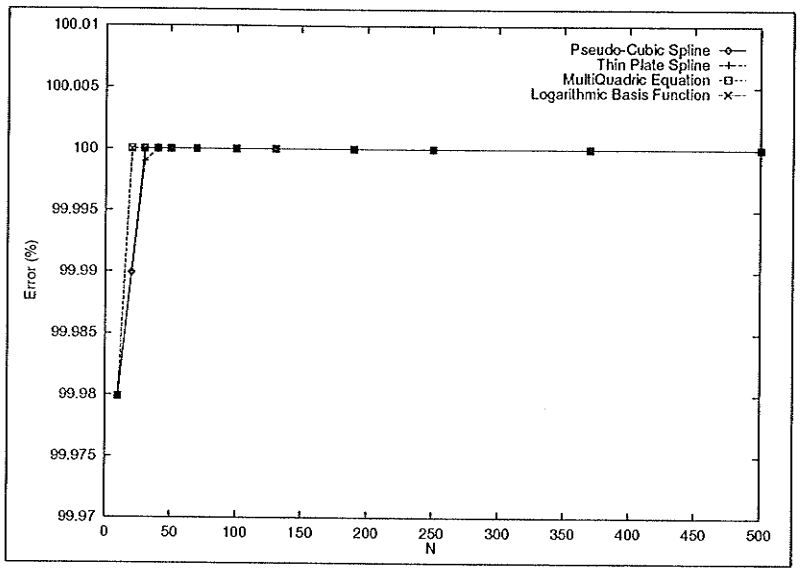

Figure 6.9: Classification results of the Vowel Experiment using Radial Kernel Classifier with four Radial Basis Functions

Recall that RKC will classify an observation $\boldsymbol{x}$ if

$$P(\omega_u)\sum_{i=1}^{N^{(u)}}\alpha_i^{(u)}\phi\left(\frac{\|\boldsymbol{x}-\boldsymbol{c}_i^{(u)}\|}{h}\right) \geq P(\omega_v)\sum_{j=1}^{N^{(v)}}\alpha_j^{(v)}\phi\left(\frac{\|\boldsymbol{x}-\boldsymbol{c}_j^{(v)}\|}{h}\right). \tag{6.13}$$

In order to classify an observation $\boldsymbol{x}$ properly, the radial kernel function of those centroids which are close to $\boldsymbol{x}$ should give a high response. Since the four RBFs studied here have a concave shape, they would provide a high response only to those centroids which are further away from $\boldsymbol{x}$. As a result, the RKC cannot classify any observation properly using concave RBFs.

6.4 Summary

From the results presented in this chapter, there are four observations.

1. Although finite support Second Order Kernels with the RKC have a faster classification speed than when the Gaussian Kernel is used, they are not able

to provide a better performance than the Gaussian Kernel.

2. The performance of the Higher-Order Kernels are generally worse than those of the Gaussian Kernel.

3. RKC loses all its discriminating power when it is used with a concave RBF.

4. The results in this chapter suggest that RKC is similar to CKC in that it requires its kernel function gives a positive response for any observation and have a convex shape. This observation is reasonable because the RKC is derived mainly from the CKC.

Chapter 7

Smoothing Parameter Selection

The smoothing parameter h is very important to the performance of the RKC. If the h is too small, the radial kernels will not be wide enough to cover all the training data. This will result in a large classification error. On the other, if the h is too large, then the boundary between classes becomes blurred. This also results in a large classification error. The selection of h can literally make or break the RKC. In this chapter, different methods for optimizing the smoothing parameter h are studied with respect to the classification error of the RKC. The process of optimizing h consists of three steps. First, one needs to specify a range or a list of h where the optimum will lie. Second, the classification error of a RKC is calculated using this list of h. Finally, the h in the list which gives the minimum classification error is selected to be the optimal h. Since the last step is trivial, this chapter focuses only on the first two steps. First, three different methods for estimating the classification errors using only the training data are studied. Next, three methods for selecting a range of h for optimizing h are studied.

7.1 Error Estimation

In an ideal situation, one would have a large number of training and testing data to calculate the classification error rate, L, for a given h. However, for most classification problems, the sample data are seldom enough even for training the classifier, let alone testing it. As a result, very often one has to estimate L using the same data used in training the classifier. Three commonly used methods for estimating L are the Resubstitution method, the Leave-One-Out method and the Bootstrap method. In this section, the error estimation performance of these three methods are compared with the minimum error, L.

The Vowel experiment was used in this section to assist the study. In this experiment, OCON and ENM were used for selecting N, the K-Means clustering was used to select the centroids, the Gaussian Kernel was used as the radial kernel function and the Euclidean metric was used.

7.1.1 Resubstitution Method

The Resubstitution method was introduced by Smith in 1947 [43]. It is the simplest and the fastest method of the three studied. The procedure of this method is as follows:

1. Train the classifier using all the training data. The training of the RKC can be further divided into the following steps:
   (a) Select N centroids using the training data.
   (b) Find the weights for each centroid by counting the training data within its cluster.
2. Classify all the training data using the trained RKC.

The resulting classification error rate, L_R, is the Resubstitution estimate of the true error L.

7.1.2 Leave-One-Out Method

The Leave-One-Out method was proposed by Lachenbruch and Mickey in 1968 [44]. This method is slower and slightly more complex than the Resubstitution method. Given n training data, the procedures for this method are as follows:

1. Select a training datum i.
2. Train the classifier with the remaining $n-1$ training data. For the RKC, this includes the following steps:
   (a) Select N centroids using the remaining $n-1$ training data.
   (b) Find the weights for each centroid by counting the number of training data within its cluster.
3. Classify the training datum i using the trained classifier.
4. Repeat step 1 to 3 until each training data has been selected once.

The resulting classification error rate, L_L, is the Leave-One-Out estimate of the true classification error L.

7.1.3 Bootstrap Method

The Bootstrap method was introduced by Efron in 1979 [45]. This method is the slowest and the most complex method of the three studied. Given a training set T with n data, the implementation of this method is as follows:

1. Set the counter *NotInBootstrap* to zero.

2. Select n data randomly from the training data with replacement to form a bootstrap set, T_B. Note that not every training datum is in the bootstrap set and the same training datum may appear more than once in the set.

3. Train the classifier with the bootstrap set, T_B. For the RKC, this includes the following steps:

   (a) Select N centroids from T_B.

   (b) Find the weights for each centroid by counting the number of data in the bootstrap set fallen within its cluster.

4. Classify the training set T using the trained classifier.

5. Count the number of training data in T that are not in the bootstrap set, T_B, and add this number to the counter *NotInBootstrap*.

6. Steps 2 to 5 are repeated B times.

7. Compute the error rate L_B by dividing the total number of classification error by the counter *NotInBootstrap*.

The resulting classification error rate, L_B, is the Bootstrap estimate of the true classification error L. In a paper by Jain and Ramaswami [46], they recommended using $B \geq 100$ for the Bootstrap method. Following their recommendation, $B = 100$ is used in the experiment.

7.1.4 Vowel Classification

This experiment is the ten-class vowel classification problem described in Section 3.2.2. To calculate the minimum error, L, the RKC was trained using the training data and the h was optimized using the test data. For the three methods studied, the training data were used for both training the RKC and for optimizing h. After the three

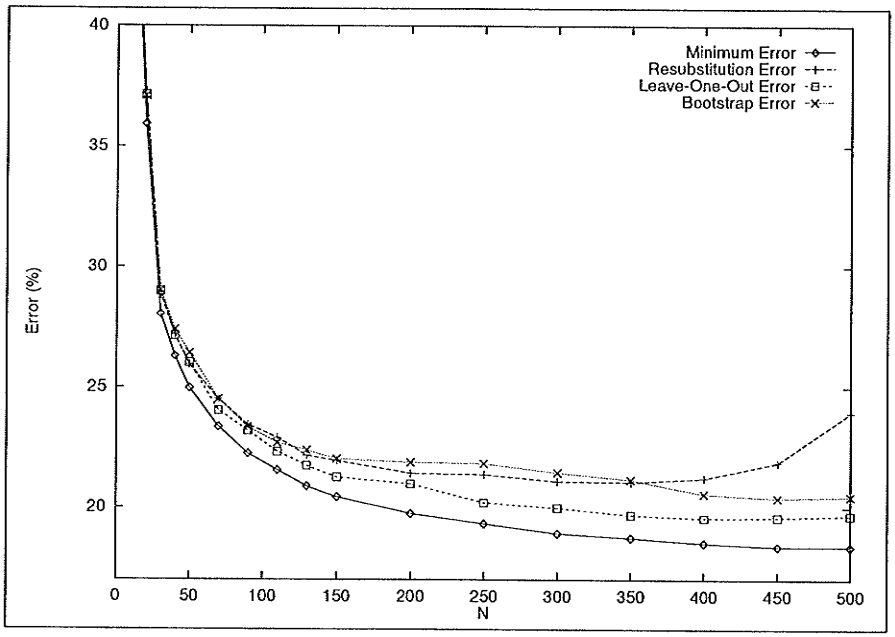

Figure 7.1: Classification results of the Vowel Experiment using three error estimation methods

methods have found the estimated optimum h, they were tested using the test data. For more details on the procedure of this experiment, please refer to Section 3.2.2.

Results and Discussion

The error rate versus N curves of the RKC are plotted in Figure 7.1 and h versus N curves are plotted in Figure 7.2. The results show that the Leave-One-Out is the best method to use in estimating the classification error of RKC. Not only did it achieve error rates closest to the minimum error L, but for most N its estimate of the optimum h were closer to the true optimum than the other two methods. In addition, the speed of the Leave-One-Out method was just slightly longer than those of the Resubstitution method. With this speed and accuracy, the Leave-One-Out method is recommended for estimating the classification error for the RKC.

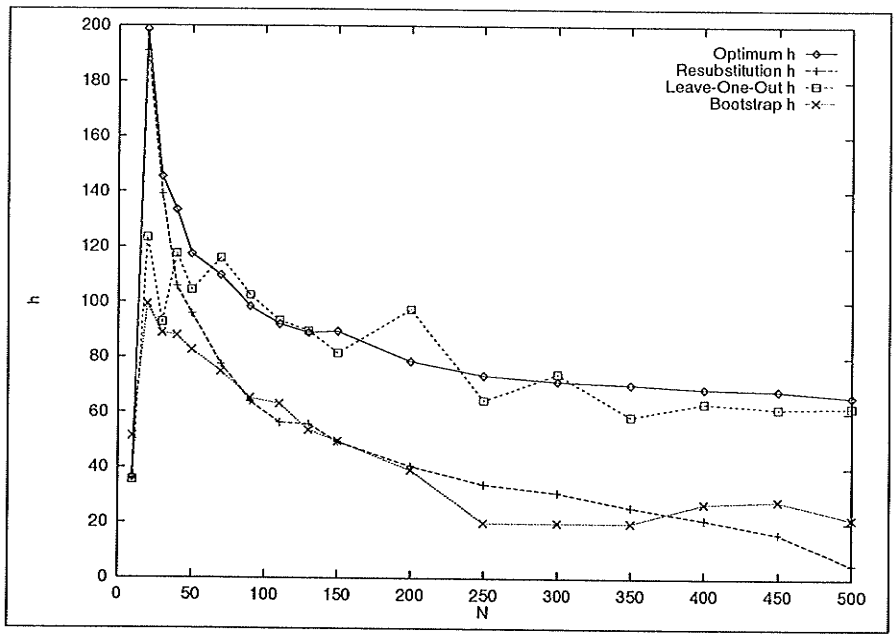

Figure 7.2: The smoothing parameter versus the number of centroids plot of the Vowel Experiment

7.2 Learning h

Once it is established what error estimation method one should use, the next step is to find a range or a list of h where the optimum h should lie. This section studies the performance of three methods for generating this list of h: the Range of h method, the Locality Index method and the Three-Point Search method.

The CKC literature only emphasizes how to estimate the classification error but not how to find the list of h for the calculation. The reason for this is probably because the optimum h for most studied problems is small — generally less than 10 — and the range where the optimum h lay can be easily guessed by experience. In the RBFN literature, only a handful of papers such as [30] by Lee, [27] by Reynolds and Tarassenko, and [47] by Hwang, Lay and Lippmann mentioned the use of a global scaling factor with the RBFN. This global scaling factor serves the same function as the smoothing parameter in CKC literature. In these three papers, only [27] had noticed the difficulties in finding the best h and proposed the Locality Index method to speed up the process of locating the optimum h. The other two papers selected h by trial and error.

There are two difficulties in finding the optimum h for RKC. The first difficulty is that without any knowledge of the range where the optimum h lies, one either has to

a. Start with a large range of h with big steps then zoom into the correct range step by step.

b. Use a small range of h hoping that the optimum h lies within it. If the optimum h is somewhere else then one will have to use another range of h and start again.

Both methods are labour intensive and time consuming. The second difficulty is that for different N the range of h is different. Thus, one will either have to guess the range of h for every N that is used or one will have to use a large range of h which would hopefully include all the optimum h. These difficulties are more serious in real

life problems because in order to classify these data, usually a large range of h is needed. For example, in the Vowel experiment described in Section 3.2.2 the range of h is between $[5, 500]$. It is these difficulties which motivates the author to propose the Three-Point Search method and to study its performance together with the other two methods.

7.2.1 Range of h (ROH)

In general, the list of h is selected based on experience or by trial and error. For most simulated problems, the optimal h usually lies between $[0.1, 5]$. If this range is used with a step size of 0.1, then one would have a list of 50 h. This method is called the Range of h method. It is commonly used both in RBFN and CKC literature.

7.2.2 Locality Index Method (LIM)

In order to increase the speed of locating the optimum h, Reynolds and Tarassenko proposed to use the Locality Index method to generate the list of h in [48]. Using this method, one will set $h = 2^{\ell}$ where ℓ is called the "locality index" and it is an integer. For example, using Locality Index, the range $[0.1 : 5]$ could be covered by using only eight ℓ ranging from -4 to 3. The price one pays for the speed of this method is the accuracy of the resulting h.

7.2.3 Three-Point Search (TPS)

The basic idea behind the Three-Point Search (TPS) is to find an estimate of the optimum h by comparing the error rates of three different h. These three h are labeled $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. If these three points form a $\vee$ shape, then the optimum h should lie between Point $\mathcal{A}$ and $\mathcal{C}$. If they form a line sloping downward to the right, or to the left, then the optimum h is beyond the range of the three points and the range

of the search have to be extended. If these three points form a horizontal line or if the range between the three points is small enough then the TPS will stop and it will label $\mathcal{B}$ as the estimated optimum h. The detailed procedure of the TPS is as follows:

1. Set point $\mathcal{A}$ and $\mathcal{C}$ to some initial values and set point $\mathcal{B}$ to the middle point between $\mathcal{A}$ and $\mathcal{C}$. In the experiment, $\mathcal{A} = 1$, $\mathcal{B} = 50.5$ and $\mathcal{C} = 100$ were used. These points are the initial list of h.

2. Set the parameters P_{Left} and P_{Right} to some initial values. The search for the optimum h will stop when

$$L_{\mathcal{A}} - L_{\mathcal{B}} \leq P_{Left}, \quad \text{and} \quad L_{\mathcal{C}} - L_{\mathcal{B}} \leq P_{Right} \tag{7.1}$$

   where $L_{\mathcal{A}}$, $L_{\mathcal{B}}$ and $L_{\mathcal{C}}$ are the classification error rates in percentage for the point $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ respectively. In the experiment, both P_{Left} and P_{Right} are set to 1%.

3. Train the RKC and then calculate the classification error in percentage using points $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. The classification error rate can be estimated using one of the three methods discussed in the Section 7.1. In the experiment, the testing data are used to calculate the classification error rate.

4. Compare the error rates $L_{\mathcal{A}}$, $L_{\mathcal{B}}$ and $L_{\mathcal{C}}$:

   (a) If $L_{\mathcal{A}} \geq L_{\mathcal{B}}$, $L_{\mathcal{C}} \geq L_{\mathcal{B}}$ and

      i. if these errors also satisfy equation (7.1) then TPS will stop and $\mathcal{B}$ will be the estimated optimum h.

      ii. if equation (7.1) is not satisfied then one will set

$$\mathcal{A} = (\mathcal{A} + \mathcal{B}) \div 2, \quad \text{and} \tag{7.2}$$

$$\mathcal{C} = (\mathcal{C} + \mathcal{B}) \div 2. \tag{7.3}$$

(b) If $L_A > L_B$ and $L_B > L_C$ then these errors form a line sloping downward to the right. In this case, the search will move to the right by setting

$$\mathcal{C} = \mathcal{C} \times 2, \quad \text{and} \tag{7.4}$$

$$\mathcal{B} = (\mathcal{A} + \mathcal{C}) \div 2. \tag{7.5}$$

(c) If $L_A < L_B$ and

i. if $L_B < L_C$, that is, the three errors form a line sloping downward to the left, then the search will move to the left by setting

$$\mathcal{A} = \mathcal{A} \div 2, \quad \text{and} \tag{7.6}$$

$$\mathcal{B} = (\mathcal{A} + \mathcal{C}) \div 2. \tag{7.7}$$

ii. if $L_C < L_A$ then these errors form the $\wedge$ shape. The search will extend to the right by setting

$$\mathcal{C} = \mathcal{C} \times 2, \quad \text{and} \tag{7.8}$$

$$\mathcal{B} = (\mathcal{A} + \mathcal{C}) \div 2. \tag{7.9}$$

iii. otherwise, the search will extend to the left by setting

$$\mathcal{A} = \mathcal{A} \div 2, \quad \text{and} \tag{7.10}$$

$$\mathcal{B} = (\mathcal{A} + \mathcal{C}) \div 2. \tag{7.11}$$

5. If the estimated optimum h is not found then go to step 3.

From experience, it usually takes TPS about 10 to 15 error calculations to find the estimated optimum h. Thus, the speed of TPS is comparable to those of the Locality Index method. The following experiment shows that the estimated optimum h obtained by the TPS is closer to those obtained by the Range of h than by the Locality Index method.

7.2.4 Vowel Classification

This experiment is the ten-class vowel classification problem described in Section 3.2.2. For the Range of h method, the range $[5, 500]$ was used with a step size of 1. For the Locality Index method, the range of ℓ from $[1, 9]$ was used with a step size 1. For the Three-Point Search method,

$$\mathcal{A} = 1, \quad \mathcal{B} = 50.5, \quad \mathcal{C} = 100, \quad \text{and} \quad P_{Left} = P_{Right} = 1\%$$

were used as the initial values. The classification error, L, was calculated using the list of h generated with the three methods, that is, a RKC was trained using the training data and its classification error was calculated using the testing data. For the procedures of this experiment, please refer to Section 3.2.2.

Results and Discussion

The error rate versus N curves of the RKC are plotted in Figure 7.3 and h versus N curves of the RKC are plotted in Figure 7.4. The results shows that the estimated classification errors and the estimated optimum h obtained by using the Three-Point Search (TPS) are closer to the results of the Range of h (ROH) than by using the Locality Index method (LIM). In other words, the TPS results are more accurate than the LIM results. This observation is not surprising since TPS is allowed to use any real h while the LIM is limited to use only a few h values. The important points about a h learning method however are not only the accuracy of this method but also on the amount of information needed to use this method and its speed of estimation.

In order to use the ROH, one needs to know the range where the optimum h lies for every N. In the Vowel experiment, the range $[5, 500]$ is selected through trial and error. This is both computationally intensive because for every N one has to calculate the classification error using a list of 495 h, and labour intensive because a wrong guess in the range of h would require a rerun of the experiment. Since one

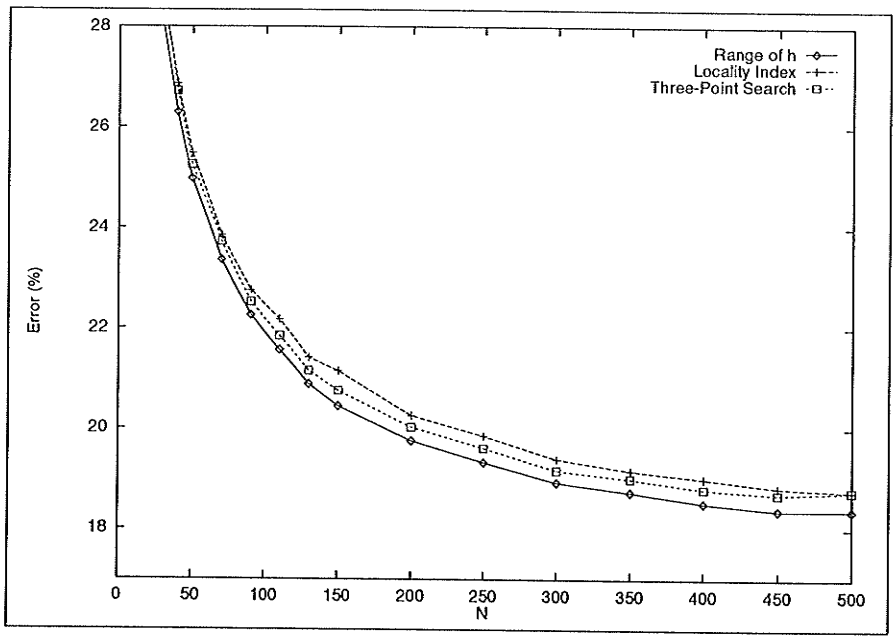

Figure 7.3: Classification Results of the Vowel Experiment using three h learning techniques

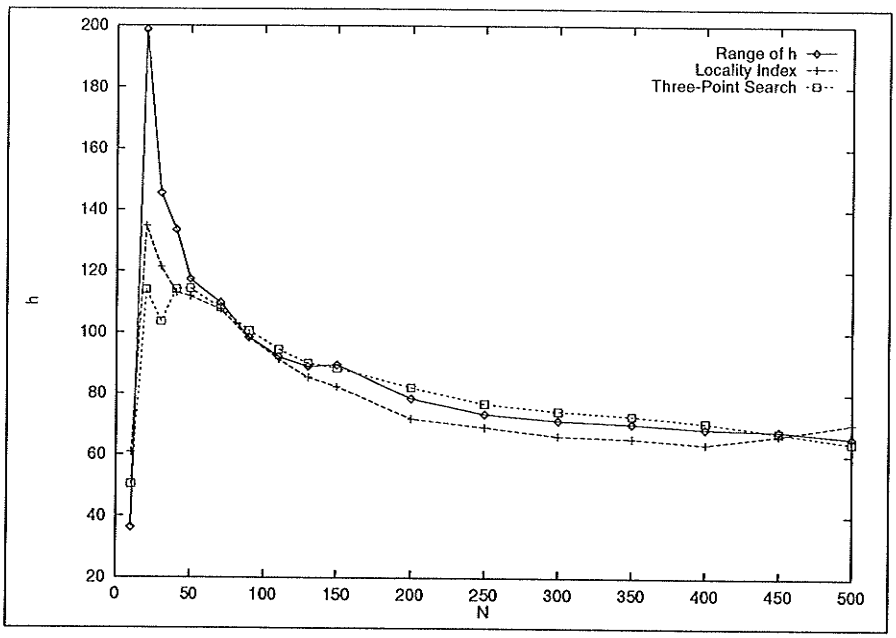

Figure 7.4: The smoothing parameter versus the number centroids plot of the Vowel Experiment

seldom has the knowledge about the range of h in advance, the use of ROH in a real life problem can be very frustrating.

The LIM is better than ROH because it only requires a few ℓ to cover a vast range of h. In the Vowel experiment, $\ell = 1, \ldots, 9$, that is a total of only 9 ℓ were used to cover the range of h from $[2, 512]$. Since LIM used a list of only 9 h, its speed is much faster than those of the ROH. The accuracy of the estimated h however is the LIM's major problem. Since ℓ can only take on integer values, the optimum h estimated by LIM is not accurate. One can relax the constraint and let ℓ take on any real number but this would turn LIM into ROH. Although LIM gains speed by using integer ℓ, it losses the accuracy in estimating h.

TPS has both the speed of LIM and the accuracy of the ROH. In the experiment, TPS required only 13 classification error calculations on average to find the estimated optimum h for a given N. Clearly the TPS is a lot faster than ROH which required 495 error calculations per N. Although the TPS required twice as many error calculations as the LIM, the speed of TPS however was just slightly slower than that of the LIM. This slight loss in speed of the TPS is offset by its adaptability and also an increase in accuracy. The adaptability of the TPS can be seen by its ability to extend its search range automatically in order to seek out the optimum h. As a result, the TPS does not require any knowledge about the range where the optimum h lies.

In a real life problem, the use of TPS is recommended to find the approximate range and an estimate of the optimum h. Then if one needs a more accurate estimate, ROH can be used to search the approximate range further.

7.3 Summary

In the first section of this chapter, it is established that the Leave-One-Out (LOO) method should be used to estimate the classification error of the RKC for its speed and accuracy. The second section shows that the Three-Point Search should be used

to estimate the optimum h because of its adaptability, speed and accuracy.

Chapter 8

Conclusions and Recommendations

8.1 Conclusions

The main goal of this thesis was to determine how the Radial Kernel Classifier (RKC) should be trained to achieve an optimum classification result. The RKC is a hybrid between the Classical Kernel Classifier (CKC) and the Radial Basis Function Networks (RBFN) which was first proposed in [2]. Since CKC and RBFN belong to two different types of Pattern Classifiers, the former belongs to the Statistical Pattern Classifiers and the latter belongs to the Neural Networks Classifiers, it was not clear from [2] how RKC should be trained for a given problem. In this thesis, the ideas in [2] are extended in order to give a recommendation as to what techniques and procedures should be used to train RKC.

By inheriting the convergence property of the CKC and the compactness of RBFN, RKC is better than both CKC and RBFN. Like the CKC, when the training data approach infinity the RKC is able to converge to the minimum error, the Bayes error. Unlike the CKC which uses all the training data for classification, however, the number of centroids used by RKC for classification is usually a lot less than the number of training data. Thus, RKC has a faster classification speed than CKC. The

use of a small number of centroids instead of the whole training set is inherited from the RBFN. Unlike the RBFN, however, RKC does not need to perform a Pseudo-Inverse on a matrix in order to calculate its weights. Instead, it finds the weights by counting the number of training data grouped within a cluster. This saves learning time and avoids the possibility of having a singular matrix when the training data is ill conditioned. These advantages of the RKC are a strong motivation to find a procedure to train it.

After reviewing all the results in the thesis, the following procedure is suggested for training RKC:

1. Based on their classes, separate the training data. This prepares for the One-Class-One-Net method of step 2.

2. Select $N^{(u)}$ centroids from the u-th class using the K-Means clustering techniques. Repeat this procedure for all S classes. The number of centroids per class should be equal for each class. Although Decision Surface Mapping (DSM) can outperform K-Means in certain problems, DSM does not give a consistent performance. Thus, the use of K-Means over DSM is preferred.

3. Set the weight, $\alpha_i^{(u)}$, of centroid $\boldsymbol{c}_i{}^{(u)}$ to the number of training data from class u grouped into its cluster $\boldsymbol{C}_i^{(u)}$.

4. Calculate the sample covariance matrix for each class. These covariance matrices are needed for the calculation of the One Class One Sigma (OCOS) Metric. If one knows that the data comes from a uniform distribution then the Euclidean metric should be used instead. Otherwise, the OCOS metric should be used.

5. Finally, optimize h by using the Three-Point Search technique for generating a list of h and the Leave-One-Out technique for estimating the classification error.

The Gaussian Kernel should be used throughout this procedure. In addition, the above procedure should be repeated for a list of N where $N = \sum_{u=1}^{S} N^{(u)}$.

The main reason why no methods which could learn the optimum N automatically were studied was because for most problems the optimum N equals to n, the number of training data. Finding this optimum N is usually not the purpose. Rather, in most problems one would like to select an N which gives an acceptable balance between classification accuracy and speed. As a result, it is better to select N based on the classification error versus N plot than to use any automated N selection technique.

8.2 Recommendations

There are two directions for possible future work. First, using the training procedures in this thesis, the classification performance of RKC could be compared with other classifiers such as the Multilayer Perceptrons and Nearest Neighbor. Second, other training techniques which could be used to further improve the performance of RKC may be studied. There are a lot of clustering techniques, smoothing parameter selection techniques and kernel functions that could be used with RKC. In this thesis, the performance of RKC was studied using only the frequently used techniques. As a result, further study in this direction is recommended.

Appendix A

Vowel Data

This is a copy of the original vowel data used in Peterson and Barney paper [33]. It was received from Richard Lippmann (rpl@sst.ll.mit.edu) through e-mail. It contains 75 of the original 76 speakers and the tokens of [AO] of three speakers are missing. This reduces the number of data points from 1520 down to 1494.

Table A.1: Vowels Label

#	Alphabet	Example	#	Alphabet	Example
1	IY	heed	6	AA	hod
2	IH	hid	7	AO	hawed
3	EH	head	8	UH	hood
4	AE	had	9	UW	who'd
5	AH	bud	10	ER	heard

Table A.2: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
2	M	IY	147	220	2220	2910	148	210	2360	3250
2	M	IH	141	410	1890	2680	139	420	1850	2500
2	M	EH	136	500	1760	2590	135	510	1710	2380
2	M	AE	128	690	1610	2560	131	700	1690	2580
2	M	AH	140	650	1080	2420	125	625	1060	2490
2	M	AA	140	650	1040	2450	136	670	1100	2430
2	M	UH	145	450	940	1910	141	410	830	2240
2	M	UW	140	280	650	3300	137	260	660	3300
2	M	ER	145	510	1210	1570	145	510	1130	1510
3	M	IY	105	250	2180	2680	111	244	2300	2780
3	M	IH	100	400	1930	2610	104	400	1990	2700
3	M	EH	100	550	1810	2500	95	540	1810	2480
3	M	AE	93	630	1710	2400	94	658	1755	2305
3	M	AH	100	600	1200	2320	105	612	1160	2350
3	M	AA	91	640	1080	2100	94	720	1090	2230
3	M	UH	114	460	1150	2290	114	456	1030	2300
3	M	UW	112	340	950	2240	112	326	900	2190
3	M	ER	100	500	1370	1780	106	530	1330	1800
4	M	IY	150	300	2240	3400	156	280	2450	3200
4	M	IH	156	450	1960	2400	146	440	2050	2360
4	M	EH	130	570	1780	2410	150	555	1890	2440
4	M	AE	125	750	1610	2340	136	770	1580	2350
4	M	AH	132	660	1200	2330	150	675	1140	2380
4	M	AA	125	750	1100	2550	138	800	1120	2500
4	M	AO	143	540	850	2320	150	555	890	2370
4	M	UH	136	460	960	2210	156	460	1000	2350
4	M	UW	140	380	950	2050	148	385	850	2330
4	M	ER	150	590	1400	1840	145	555	1430	1730
5	M	IY	140	310	2310	2820	131	260	2250	2850
5	M	IH	137	440	2060	2640	134	430	1880	2450
5	M	EH	140	580	1910	2500	137	550	1770	2400
5	M	AE	143	830	1720	2180	135	750	1690	2320
5	M	AH	136	630	1300	1950	130	650	1170	2000
5	M	AA	131	760	1220	2140	126	720	1260	2020
5	M	AO	136	540	970	1980	124	550	880	1950

Table A.3: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
5	M	UH	133	470	1040	1990	132	490	990	1920
5	M	UW	141	380	950	2140	133	330	800	2130
5	M	ER	143	560	1510	1800	136	510	1460	1700
6	M	IY	125	312	2350	2800	119	330	2430	2870
6	M	IH	133	420	2000	2660	125	313	2000	2750
6	M	EH	120	600	1860	2500	114	570	1830	2570
6	M	AE	119	676	1670	2540	125	725	1687	2500
6	M	AH	118	680	1150	2560	125	726	1270	2560
6	M	AA	125	740	1100	2680	113	670	960	2650
6	M	AO	120	660	1030	2690	125	720	960	2700
6	M	UH	120	456	1080	2520	120	450	1140	2600
6	M	UW	120	313	830	2300	125	288	938	2450
6	M	ER	120	503	1305	1775	120	505	1320	1750
7	M	IY	186	320	2320	3120	172	310	2280	3020
7	M	IH	167	470	2000	2660	170	410	2040	2715
7	M	EH	167	630	1900	2860	146	614	1840	2770
7	M	AE	143	740	1800	2450	162	775	1810	2200
7	M	AH	167	620	1240	2410	160	640	1250	2400
7	M	AA	162	650	970	2580	163	650	980	2350
7	M	AO	145	430	720	2450	171	510	800	2500
7	M	UH	170	460	1120	2150	170	493	1120	2300
7	M	UW	175	380	1040	2260	200	400	1000	2350
7	M	ER	167	570	1300	1750	157	565	1370	1710
8	M	IY	105	230	2480	3200	109	218	2380	3100
8	M	IH	110	320	2200	2680	103	206	2130	2570
8	M	EH	107	430	2100	2630	105	515	1760	2470
8	M	AE	107	514	2060	2600	106	552	1820	2500
8	M	AH	108	640	1300	2300	104	624	1350	2410
8	M	AA	111	714	1170	2420	97	650	1150	2350
8	M	AO	107	590	965	2500	109	578	970	2460
8	M	UH	111	467	1110	2400	105	475	1220	2310
8	M	UW	107	270	910	2200	108	260	975	2320
8	M	ER	107	460	1400	1790	103	425	1410	1760
9	M	IY	175	316	2200	2800	175	280	2275	2775
9	M	IH	167	450	1820	2475	167	434	1850	2425

Table A.4: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
9	M	EH	157	582	1725	2375	158	586	1800	2425
9	M	AE	150	600	1750	2375	145	582	1775	2375
9	M	AH	145	626	1125	2200	160	641	1120	2225
9	M	AA	144	708	1054	2420	150	705	1050	2375
9	M	AO	146	614	848	2200	143	600	860	2175
9	M	UH	167	500	1000	2325	167	500	1000	2325
9	M	UW	167	334	1150	2200	183	312	1020	2300
9	M	ER	157	518	1305	1570	157	504	1210	1510
10	M	IY	129	260	2260	2820	125	250	2200	2825
10	M	IH	146	400	2040	2500	144	389	2000	2425
10	M	EH	126	500	1870	2500	125	500	1775	2400
10	M	AE	110	660	1650	2500	120	624	1700	2475
10	M	AH	122	650	1220	2550	120	672	1260	2500
10	M	AA	114	750	1080	2680	114	777	1026	2625
10	M	AO	115	580	800	2650	117	585	819	2625
10	M	UH	140	480	950	2500	127	461	993	2350
10	M	UW	140	280	950	2300	133	266	920	2300
10	M	ER	128	500	1340	1700	133	532	1275	1600
11	M	IY	146	248	2225	3100	140	238	2175	3075
11	M	IH	150	405	1925	2550	138	416	1940	2600
11	M	EH	147	588	1790	2500	133	586	1725	2650
11	M	AE	145	725	1700	2425	127	710	1650	2220
11	M	AH	136	586	1078	2300	136	627	1038	2360
11	M	AA	145	725	1046	2325	131	746	1018	2300
11	M	AO	140	560	840	2500	140	560	924	2350
11	M	UH	150	495	1080	2275	143	430	1030	2275
11	M	UW	162	290	760	2300	157	315	850	2025
11	M	ER	150	511	1561	1876	138	530	1450	1887
12	M	IY	110	220	2410	3000	125	240	2440	3280
12	M	IH	120	450	1880	2450	118	380	1930	2420
12	M	EH	115	560	1650	2300	123	560	1720	2300
12	M	AE	110	680	1720	2330	133	630	1680	2280
12	M	AH	110	560	1430	2250	120	560	1390	2240
12	M	AA	108	800	1330	2260	110	740	1240	2280
12	M	AO	120	600	920	2080	133	580	910	2000

Table A.5: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
12	M	UH	130	400	1200	2210	110	420	1230	2230
12	M	UW	122	300	900	2130	123	260	1010	2240
12	M	ER	125	400	1450	1650	128	360	1410	1640
13	M	IY	142	290	2290	2600	135	260	2290	2700
13	M	IH	132	390	1950	2550	135	400	1900	2450
13	M	EH	124	490	1740	2500	125	500	1780	2430
13	M	AE	125	660	1630	2500	132	670	1630	2380
13	M	AH	140	600	1220	2530	125	600	1210	2430
13	M	AA	125	680	1120	2630	128	670	1100	2700
13	M	AO	127	510	720	2450	120	480	710	2540
13	M	UH	133	380	910	2350	140	440	1030	2400
13	M	UW	127	350	720	2750	140	380	740	2880
13	M	ER	128	430	1370	1610	135	440	1360	1600
14	M	IY	114	228	2350	2860	118	220	2350	2920
14	M	IH	110	407	2070	2500	112	420	1900	2450
14	M	EH	106	445	2020	2420	115	470	2020	2500
14	M	AE	103	721	1680	2400	109	750	1710	2440
14	M	AH	104	552	1122	2500	115	580	1150	2600
14	M	AA	98	686	1078	2570	103	700	1050	2680
14	M	AO	102	560	665	2620	106	550	650	2700
14	M	UH	112	448	980	2370	104	410	940	2370
14	M	UW	116	232	696	2200	117	222	665	2080
14	M	ER	120	432	1300	1400	111	420	1300	1570
15	M	IY	121	230	2100	2850	118	240	2000	2980
15	M	IH	130	365	1900	2340	119	300	2040	2560
15	M	EH	112	440	1980	2310	120	410	2050	2500
15	M	AE	133	620	1710	2110	124	660	1800	2150
15	M	AH	120	660	1000	2380	110	660	960	2450
15	M	AA	122	600	830	2250	119	620	820	2400
15	M	AO	117	500	620	2250	106	550	700	2550
15	M	UH	140	390	730	2180	130	360	740	2200
15	M	UW	131	260	720	2100	132	260	740	2040
15	M	ER	125	450	1230	1600	127	460	1300	1650
16	M	IY	150	300	2355	3250	150	300	2460	3280
16	M	IH	160	385	2242	2805	150	407	2250	2780

Table A.6: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
16	M	EH	140	504	2090	2720	146	543	1980	2640
16	M	AE	133	680	1958	2542	141	708	1840	2535
16	M	AH	150	675	1320	2550	150	704	1393	2550
16	M	AA	137	825	1168	2750	135	840	1210	2680
16	M	AO	143	671	1000	2670	147	690	968	2660
16	M	UH	143	443	1273	2430	153	459	1286	2410
16	M	UW	146	395	1300	2160	153	400	1320	2150
16	M	ER	140	532	1500	1890	146	538	1460	1818
17	M	IY	120	264	2290	2700	128	256	2305	2635
17	M	IH	112	380	1880	2440	115	346	1930	2390
17	M	EH	100	510	1780	2300	108	520	1730	2275
17	M	AE	100	630	1770	2350	105	630	1642	2170
17	M	AH	103	601	1273	2130	105	590	1283	2150
17	M	AA	100	750	1150	2440	95	703	1092	2320
17	M	AO	97	565	780	2350	106	584	849	2460
17	M	UH	105	420	1100	2140	111	422	1200	2175
17	M	UW	117	315	1080	2260	125	326	1125	2210
17	M	ER	111	444	1300	1625	109	469	1288	1600
18	M	IY	124	210	2100	3090	130	220	2080	3180
18	M	IH	128	280	2000	2710	130	310	1950	2670
18	M	EH	121	470	1910	2580	129	490	1930	2650
18	M	AE	116	640	1620	2200	118	650	1580	2360
18	M	AH	121	610	1100	2230	126	620	1120	2330
18	M	AA	118	700	1100	2240	120	670	1100	2220
18	M	AO	122	460	720	2180	118	470	690	2200
18	M	UH	129	320	770	1860	130	310	790	1920
18	M	UW	140	210	670	1900	148	240	730	1850
18	M	ER	128	390	1320	1550	124	420	1240	1510
19	M	IY	129	190	2650	3280	135	190	2700	3170
19	M	IH	132	370	1750	2700	130	370	1800	2750
19	M	EH	122	370	1680	2560	125	375	1700	2500
19	M	AE	121	550	1570	2600	120	530	1610	2650
19	M	AH	118	570	1050	2500	125	590	1100	2480
19	M	AA	112	640	970	2870	122	670	980	2900
19	M	AO	113	560	860	2900	121	570	820	2820

Table A.7: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
19	M	UH	125	350	1000	2500	130	380	920	2370
19	M	UW	130	250	1000	2100	140	210	960	1940
19	M	ER	130	360	1300	1920	133	370	1300	1760
20	M	IY	127	250	2180	2660	131	260	2210	2780
20	M	IH	121	400	1900	2440	122	350	1980	2480
20	M	EH	116	560	1670	2310	124	530	1700	2380
20	M	AE	120	680	1470	2280	119	620	1580	2320
20	M	AH	120	620	1100	2390	125	640	1110	2370
20	M	AA	115	630	980	2330	121	670	940	2380
20	M	AO	112	560	790	2480	120	610	840	2420
20	M	UH	121	360	860	2200	120	400	840	2200
20	M	UW	140	280	670	2140	126	250	720	2190
20	M	ER	120	480	1410	1760	121	470	1330	1700
21	M	IY	155	280	2400	2910	150	300	2320	2960
21	M	IH	142	410	2060	2680	150	450	2050	2670
21	M	EH	135	540	1900	2530	135	540	1920	2520
21	M	AE	138	620	1800	2440	140	690	1820	2480
21	M	AH	150	630	1200	2600	140	680	1290	2600
21	M	AA	145	740	1110	2500	143	700	1060	2720
21	M	AO	146	600	970	2570	138	650	880	2660
21	M	UH	142	430	1130	2440	143	430	1150	2420
21	M	UW	142	280	990	2330	145	290	1000	2300
21	M	ER	150	420	1350	1600	150	450	1350	1600
22	M	IY	135	300	2300	2800	135	350	2240	2760
22	M	IH	136	410	2200	2680	138	440	2080	2520
22	M	EH	133	580	1870	2320	127	520	1900	2400
22	M	AE	130	760	1920	2480	132	670	1850	2560
22	M	AH	139	810	1110	2100	131	770	1150	2100
22	M	AA	141	700	1040	2120	125	750	1160	2080
22	M	AO	133	670	920	2240	142	570	850	2250
22	M	UH	140	550	970	2200	141	490	870	2240
22	M	UW	150	300	600	2300	148	230	570	2100
22	M	ER	140	560	1520	2100	140	540	1570	2050
23	M	IY	125	240	2100	2900	119	240	2150	2860
23	M	IH	130	380	1870	2450	120	430	1710	2350

Table A.8: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
23	M	EH	119	580	1770	2500	117	570	1750	2400
23	M	AE	115	760	1580	2440	110	715	1500	2300
23	M	AH	124	620	880	2500	124	650	1000	2520
23	M	AA	119	710	950	2520	120	690	960	2520
23	M	AO	125	460	610	2500	120	470	710	2500
23	M	UH	125	390	900	2100	125	460	920	2140
23	M	UW	125	250	690	2080	130	270	650	2050
23	M	ER	122	540	1280	1720	118	510	1280	1650
24	M	IY	148	280	2450	2700	160	288	2500	2880
24	M	IH	160	400	2080	2530	153	384	2110	2500
24	M	EH	138	590	1900	2200	153	583	1840	2250
24	M	AE	145	680	1850	2400	140	685	1780	2160
24	M	AH	143	660	1370	2110	145	680	1300	2100
24	M	AA	140	760	1260	2120	135	770	1140	2020
24	M	AO	145	500	800	1850	132	600	1000	2000
24	M	UH	157	380	1060	1950	150	470	1220	2150
24	M	UW	162	324	800	2220	139	290	800	2150
24	M	ER	150	560	1350	1780	150	600	1470	1820
25	M	IY	110	250	2190	3000	106	254	2085	2890
25	M	IH	111	330	1967	2670	108	430	1940	2590
25	M	EH	116	464	2100	2700	105	504	1995	2780
25	M	AE	94	595	1900	2700	100	670	1860	2500
25	M	AH	96	620	1200	2420	105	630	1127	2420
25	M	AA	100	750	1160	2360	96	740	1155	2330
25	M	AO	101	460	740	2300	105	494	789	2420
25	M	UH	113	400	1020	2200	128	450	1028	2160
25	M	UW	140	392	1000	2120	116	350	898	2140
25	M	ER	117	547	1340	1688	128	512	1280	1570
26	M	IY	123	246	2185	2730	133	267	2280	2800
26	M	IH	140	420	2300	2800	120	384	2110	2620
26	M	EH	120	480	1920	2540	112	551	1788	2450
26	M	AE	114	628	1837	2570	111	622	1890	2560
26	M	AH	114	628	1254	2470	114	617	1255	2480
26	M	AA	117	690	1072	2660	103	630	1000	2530
26	M	AO	117	510	700	2650	120	504	756	2540

Table A.9: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
26	M	UH	122	465	990	2440	125	462	976	2450
26	M	UW	120	324	708	2440	157	387	786	2518
26	M	ER	122	488	1468	1712	118	472	1465	1725
27	M	IY	138	275	2060	2800	136	270	2020	2790
27	M	IH	133	349	2030	2760	136	340	1940	2560
27	M	EH	120	444	1800	2500	127	380	1800	2440
27	M	AE	125	688	1600	2300	122	660	1570	2380
27	M	AH	128	565	1157	2310	130	550	1150	2250
27	M	AA	125	712	1024	2250	125	670	1080	2300
27	M	AO	125	550	913	2360	126	550	890	2280
27	M	UH	128	360	1028	2160	140	390	1060	2150
27	M	UW	133	294	930	2050	140	280	1000	2160
27	M	ER	125	440	1250	1625	130	480	1160	1520
28	M	IY	125	320	2160	2900	133	267	2230	3000
28	M	IH	115	440	1750	2400	116	390	1780	2450
28	M	EH	117	525	1800	2480	110	520	1750	2390
28	M	AE	111	660	1600	2400	120	720	1680	2430
28	M	AH	117	600	1250	2300	125	575	1170	2240
28	M	AA	111	730	1160	2340	117	860	1280	2470
28	M	AO	114	560	810	2290	116	584	840	2280
28	M	UH	130	455	970	2140	120	456	1040	2038
28	M	UW	125	350	820	2130	128	366	772	2058
28	M	ER	111	450	1420	1870	118	472	1430	1840
29	M	IY	133	333	2305	3200	131	326	2260	3030
29	M	IH	125	375	2188	2750	133	400	2150	2680
29	M	EH	125	500	1980	2480	150	480	1950	2340
29	M	AE	116	640	1710	2450	123	615	1720	2220
29	M	AH	116	583	1110	2360	117	608	1120	2700
29	M	AA	111	777	1170	2600	114	750	1175	2820
29	M	AO	105	630	891	2519	114	572	924	2660
29	M	UH	125	438	975	2300	140	420	938	2300
29	M	UW	133	333	800	2130	140	320	840	2150
29	M	ER	120	480	1320	1870	127	483	1335	1844
30	M	IY	166	267	2300	2940	156	220	2300	2900
30	M	IH	154	431	2040	2460	155	360	2010	2400

Table A.10: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
30	M	EH	150	565	1950	2500	180	540	2000	2450
30	M	AE	143	600	2000	2570	138	590	1950	2460
30	M	AH	157	630	1140	2200	186	630	1170	2280
30	M	AA	146	730	1048	2450	155	730	1130	2320
30	M	AO	150	600	900	2400	178	640	890	2280
30	M	UH	160	448	960	2200	196	450	1000	2180
30	M	UW	167	333	835	2170	198	280	750	2170
30	M	ER	163	488	1300	1600	163	490	1380	1620
31	M	IY	120	312	2380	2900	120	300	2350	3000
31	M	IH	140	490	2000	2620	140	490	1960	2600
31	M	EH	125	640	2000	2620	111	555	1870	2540
31	M	AE	112	697	1610	2540	114	684	1634	2510
31	M	AH	115	633	1260	2530	120	660	1213	2460
31	M	AA	112	730	1203	2700	107	752	1125	2620
31	M	AO	108	507	755	2420	116	538	816	2450
31	M	UH	114	456	1040	2300	120	480	1120	2160
31	M	UW	123	344	960	2150	125	350	1000	2250
31	M	ER	112	539	1370	1800	117	549	1353	1728
32	M	IY	146	292	2500	3150	133	266	2370	3100
32	M	IH	143	372	2220	2640	131	350	2130	2610
32	M	EH	133	574	1840	2260	133	563	1960	2450
32	M	AE	125	650	1738	2400	130	663	1820	2400
32	M	AH	137	600	1370	2180	125	625	1312	2250
32	M	AA	133	735	1070	2100	117	713	1180	2200
32	M	AO	125	625	875	2180	115	700	1000	2250
32	M	UH	150	420	1100	2000	140	420	1120	2100
32	M	UW	125	350	980	2200	133	320	918	2100
32	M	ER	143	554	1480	1800	128	484	1505	1890
33	M	IY	143	286	2415	2860	150	300	2415	2860
33	M	IH	140	400	1980	2500	145	407	2095	2620
33	M	EH	125	525	1988	2610	144	553	1935	2530
33	M	AE	133	640	1773	2490	133	640	1840	2560
33	M	AH	143	672	1272	2640	146	658	1241	2560
33	M	AA	130	780	1170	2640	131	788	1115	2645
33	M	AO	138	633	891	2500	150	600	935	2550

Table A.11: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
33	M	UH	175	490	1102	2420	154	492	1077	2306
33	M	UW	160	320	960	2240	160	320	960	2290
33	M	ER	143	543	1310	1643	145	508	1309	1600
34	F	IY	230	370	2670	3100	234	390	2760	3060
34	F	IH	234	468	2330	2930	205	410	2380	2950
34	F	EH	190	550	2200	2880	191	570	2100	3040
34	F	AE	200	800	1980	2810	192	860	1920	2850
34	F	AH	227	635	1200	3250	200	700	1200	3100
34	F	AA	210	880	1240	2870	188	830	1200	2880
34	F	AO	207	570	830	3300	200	700	1000	3130
34	F	UH	240	410	940	3040	225	450	970	3190
34	F	UW	238	480	955	2960	208	395	810	2900
34	F	ER	200	500	1850	2100	200	560	1750	2100
35	F	IY	225	270	2760	3550	240	290	2700	3350
35	F	IH	245	460	2500	3220	220	410	2400	3240
35	F	EH	220	620	2300	3200	210	630	2300	3170
35	F	AE	220	820	2180	2850	195	740	2120	3070
35	F	AH	240	800	1300	2900	225	760	1400	2830
35	F	AA	214	850	1120	2620	190	880	1220	2850
35	F	AO	228	460	900	2830	222	440	880	2850
35	F	UH	250	500	1040	2750	245	490	1000	2720
35	F	UW	250	400	940	2720	245	410	860	2700
35	F	ER	225	440	1560	1750	210	420	1600	1750
36	F	IY	210	290	2700	3020	215	280	2630	3240
36	F	IH	211	420	2300	2950	211	420	2220	2980
36	F	EH	207	640	2120	2900	221	700	2000	2900
36	F	AE	212	1000	1830	2820	204	980	1800	2820
36	F	AH	205	780	1410	2720	208	710	1450	2750
36	F	AA	205	950	1280	2600	210	870	1260	2740
36	F	AO	203	610	900	2710	210	630	840	2700
36	F	UH	211	440	1050	2780	210	420	1050	2740
36	F	UW	222	380	860	2500	208	330	750	2740
36	F	ER	208	580	1450	1720	212	540	1560	1900
37	F	IY	210	294	2800	3100	222	270	2880	3160
37	F	IH	202	420	2430	3030	212	420	2370	2930

Table A.12: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
37	F	EH	200	580	2180	2770	217	540	2160	2770
37	F	AE	200	820	1970	2620	210	840	2000	2700
37	F	AH	208	690	1200	2900	201	666	1206	2900
37	F	AA	200	800	1200	2920	190	760	1140	2850
37	F	AO	200	560	760	2800	207	560	770	3000
37	F	UH	215	430	1075	2580	213	430	1000	2700
37	F	UW	220	330	840	2550	213	280	850	2500
37	F	ER	205	430	1800	1930	200	420	1740	1960
38	F	IY	175	350	2800	3160	187	338	2870	3300
38	F	IH	200	400	2540	3200	210	420	2680	3000
38	F	EH	180	518	2470	3200	200	600	2400	3150
38	F	AE	171	773	2000	2870	175	875	2100	2970
38	F	AH	183	733	1468	2700	200	740	1280	2900
38	F	AA	178	730	1210	2740	175	735	1220	2850
38	F	AO	160	560	960	2850	192	536	850	2850
38	F	UH	212	424	1040	2780	200	520	1060	2670
38	F	UW	190	380	770	2900	187	340	750	2780
38	F	ER	177	490	2120	2480	197	493	1930	2300
39	F	IY	250	325	2700	3100	225	310	2750	3225
39	F	IH	214	350	2580	3000	267	390	2700	3200
39	F	EH	233	560	2330	2800	200	520	2500	3000
39	F	AE	171	806	1970	2600	150	825	1860	2550
39	F	AH	186	708	1485	2760	188	676	1500	2590
39	F	AA	200	800	1200	2800	205	714	1154	2850
39	F	AO	267	530	800	2780	180	485	810	2750
39	F	UH	214	450	1460	2550	233	467	1400	2450
39	F	UW	225	450	1080	2350	200	400	1000	2400
39	F	ER	193	524	1700	2130	180	507	1800	2380
40	F	IY	200	300	3100	3400	216	300	3100	3500
40	F	IH	214	428	2570	3000	220	440	2640	3080
40	F	EH	210	528	2540	3170	210	504	2520	3200
40	F	AE	187	940	2250	2760	200	820	2200	2920
40	F	AH	204	816	1450	2700	214	858	1500	2700
40	F	AA	200	960	1280	3000	180	1040	1300	3000
40	F	AO	220	520	880	2500	217	574	890	2510

Table A.13: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
40	F	UH	233	466	1330	2750	233	466	1165	2800
40	F	UW	180	300	850	2800	175	350	840	2750
40	F	ER	216	432	1790	2060	219	360	1900	2320
41	F	IY	225	337	2700	3300	233	340	2720	3200
41	F	IH	237	474	2370	3095	237	475	2400	3090
41	F	EH	229	526	2360	3090	233	580	2360	3150
41	F	AE	230	690	2185	2990	220	660	2200	3020
41	F	AH	225	675	1551	2923	233	690	1630	2900
41	F	AA	222	845	1334	2890	233	888	1290	2800
41	F	AO	225	631	923	2250	233	543	980	2300
41	F	UH	233	537	1360	2920	240	480	1345	2680
41	F	UW	235	400	1180	2760	233	396	1120	2560
41	F	ER	225	450	1640	2250	233	489	1630	2090
42	F	IY	225	225	2760	3900	230	230	2850	3800
42	F	IH	238	429	2560	3200	230	430	2575	3100
42	F	EH	214	579	2570	3300	214	536	2570	3100
42	F	AE	205	823	2220	2870	200	800	2100	2900
42	F	AH	250	750	1500	2750	217	738	1300	2820
42	F	AA	200	840	1300	3100	206	990	1340	3100
42	F	AO	214	579	856	2790	205	545	905	2750
42	F	UH	233	490	1220	2610	250	513	1500	2650
42	F	UW	250	400	1250	2500	225	405	1080	2500
42	F	ER	233	466	1860	2260	225	540	1780	2220
43	F	IY	240	290	3000	3840	250	325	2900	3500
43	F	IH	250	500	2370	3120	238	476	2380	3090
43	F	EH	238	760	2380	3205	233	746	2290	3030
43	F	AE	206	1008	1990	2870	200	1040	2000	2800
43	F	AH	220	830	1540	2860	237	900	1510	2840
43	F	AA	206	970	1343	3018	236	592	1230	2600
43	F	AO	233	650	900	2920	229	687	1060	2780
43	F	UH	233	512	1211	2630	233	467	1167	2595
43	F	UW	250	450	875	2750	233	420	935	2710
43	F	ER	230	622	1750	2070	225	652	1710	2043
44	F	IY	255	275	2800	3310	245	245	2800	3300
44	F	IH	267	534	2500	3250	264	528	2640	3370

Table A.14: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
44	F	EH	238	700	2380	3250	250	750	2480	3000
44	F	AE	237	1020	1900	2960	233	1005	2050	2870
44	F	AH	263	750	1500	2850	250	850	1400	2750
44	F	AA	258	978	1290	2840	246	935	1230	2730
44	F	AO	250	500	750	2750	243	632	850	2850
44	F	UH	250	350	1170	2750	266	450	1000	2800
44	F	UW	256	358	640	2560	250	300	750	2500
44	F	ER	260	520	1560	1820	250	500	1500	1750
45	F	IY	236	236	2790	3760	242	242	2770	3800
45	F	IH	222	444	2555	3110	242	420	2700	3120
45	F	EH	226	634	2325	2940	225	608	2475	3100
45	F	AE	210	1010	2060	2900	200	980	2160	2920
45	F	AH	217	818	1450	2500	200	750	1280	2650
45	F	AA	220	820	1200	2640	210	900	1120	2900
45	F	AO	220	440	749	2640	210	567	752	2600
45	F	UH	204	460	1045	2504	240	480	1105	2400
45	F	UW	250	420	1000	2500	275	350	1100	2400
45	F	ER	217	487	1500	1780	206	467	1420	1640
46	F	IY	225	360	2920	3400	233	340	2840	3300
46	F	IH	257	514	2570	3070	238	500	2680	3260
46	F	EH	238	650	2495	3090	216	650	2380	3030
46	F	AE	225	1020	2030	2700	225	1000	2200	2770
46	F	AH	225	788	1462	2920	217	736	1500	2900
46	F	AA	214	987	1330	2830	214	1009	1415	3080
46	F	AO	226	672	1084	2495	209	627	1045	2504
46	F	UH	250	500	1200	2450	230	460	1150	2880
46	F	UW	267	420	990	2860	190	380	893	2920
46	F	ER	246	610	1630	2020	225	585	1700	1850
47	F	IY	285	285	2900	3500	286	310	2900	3400
47	F	IH	297	480	2670	3260	220	440	2620	3380
47	F	EH	173	550	2370	3140	260	520	2340	3040
47	F	AE	167	790	2180	3020	280	840	2160	3020
47	F	AH	280	840	1400	2750	270	760	1330	2950
47	F	AA	252	900	1290	2750	260	900	1240	3110
47	F	AO	175	700	1050	2750	190	720	1080	3030

Table A.15: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
47	F	UH	286	540	1200	2860	205	570	1200	2970
47	F	UW	328	400	980	2630	290	440	990	2900
47	F	ER	286	570	2000	2480	260	510	1850	2350
48	F	IY	170	340	2750	3120	238	360	2760	3120
48	F	IH	167	480	2390	2950	194	520	2450	3000
48	F	EH	220	620	2520	2920	222	620	2440	2880
48	F	AE	222	1110	2160	2700	214	1070	1960	2750
48	F	AH	217	820	1240	2600	216	860	1300	2670
48	F	AA	150	840	1110	2930	170	850	1120	2850
48	F	AO	200	500	700	2930	212	380	720	2700
48	F	UH	235	400	940	2820	214	380	860	2680
48	F	UW	196	330	760	2870	188	350	710	2760
48	F	ER	182	550	1780	2080	201	600	1750	2000
49	F	IY	200	320	2360	2980	203	304	2380	3050
49	F	IH	211	444	2220	2740	210	420	2090	2780
49	F	EH	200	500	2350	2830	200	600	2200	2700
49	F	AE	192	845	1700	2300	187	860	1724	2530
49	F	AH	200	720	1440	2380	191	707	1470	2440
49	F	AA	200	700	1080	2420	192	767	1150	2590
49	F	AO	200	600	860	2410	200	600	900	2400
49	F	UH	210	546	1090	2400	210	462	1240	2310
49	F	UW	257	360	930	2260	220	440	1100	2300
49	F	ER	200	540	1400	1800	204	460	1350	1560
50	F	IY	203	406	2600	2945	200	400	2600	3100
50	F	IH	200	460	2300	2800	210	420	2305	2835
50	F	EH	190	570	2100	2720	207	538	2175	2880
50	F	AE	189	850	1853	2685	193	830	1800	2620
50	F	AH	200	720	1500	2560	200	800	1400	2420
50	F	AA	194	915	1280	2530	206	723	1196	2600
50	F	AO	192	575	1073	2490	200	600	1100	2600
50	F	UH	202	520	1210	2420	212	468	1275	2550
50	F	UW	207	370	1000	2470	205	330	970	2460
50	F	ER	200	560	1600	1900	206	514	1540	1955
51	F	IY	240	380	2880	3360	250	380	2820	3300
51	F	IH	233	514	2600	2930	237	473	2660	2970

Table A.16: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
51	F	EH	223	567	2460	3122	224	521	2460	2920
51	F	AE	218	808	2070	2880	203	678	2420	3080
51	F	AH	200	800	1340	2700	214	772	1280	2660
51	F	AA	183	843	1190	2860	205	740	1160	2780
51	F	AO	222	623	1022	2700	220	594	990	2640
51	F	UH	240	480	960	2820	242	484	900	2640
51	F	UW	233	370	933	2520	250	325	750	2500
51	F	ER	225	450	1680	2050	233	466	1630	1865
52	F	IY	200	320	2750	3100	178	356	2755	3200
52	F	IH	194	388	2622	3050	194	426	2460	3040
52	F	EH	187	592	2242	2765	191	535	2290	2870
52	F	AE	188	750	2060	2770	162	650	2110	2618
52	F	AH	187	618	1518	2700	183	624	1430	2660
52	F	AA	163	766	1180	2340	167	750	1065	2640
52	F	AO	170	595	918	2600	176	630	985	2630
52	F	UH	200	420	1200	2600	200	460	1260	2640
52	F	UW	187	375	1124	2685	188	375	1143	2700
52	F	ER	180	504	1565	1835	183	513	1578	1830
53	F	IY	280	357	2800	3360	275	340	2860	3350
53	F	IH	290	480	2600	3060	292	465	2598	3060
53	F	EH	250	700	2350	2980	240	737	2325	3100
53	F	AE	200	960	2100	3000	217	1030	2200	3260
53	F	AH	275	920	1512	2950	260	910	1688	3050
53	F	AA	275	990	1237	2360	267	987	1172	3180
53	F	AO	267	587	1068	3270	293	560	990	3150
53	F	UH	275	520	1350	3190	280	510	1415	3130
53	F	UW	300	420	1045	3060	300	390	960	3030
53	F	ER	230	460	1860	2250	214	504	1820	2290
54	F	IY	200	240	2760	3700	220	220	2850	3800
54	F	IH	228	319	2500	3020	216	324	2500	3010
54	F	EH	220	616	2380	2900	212	615	2300	2800
54	F	AE	212	710	2120	2600	210	690	2250	2680
54	F	AH	221	800	1520	2380	210	780	1470	2400
54	F	AA	199	995	1392	2290	200	1000	1400	2440
54	F	AO	205	656	944	2250	200	720	960	2380

Table A.17: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
54	F	UH	223	335	1049	2470	210	420	1009	2300
54	F	UW	219	329	877	2550	230	340	900	2530
54	F	ER	206	400	1380	1560	201	400	1240	1480
55	F	IY	220	286	2800	3550	241	289	2800	3400
55	F	IH	225	383	2420	3080	240	384	2400	3050
55	F	EH	209	418	2430	3110	230	460	2300	3050
55	F	AE	187	861	2100	2800	224	896	2040	3000
55	F	AH	218	654	1160	2800	230	690	1195	2770
55	F	AA	208	860	1103	2700	212	806	1060	2850
55	F	AO	202	606	910	2900	201	583	860	2840
55	F	UH	225	340	900	2650	235	470	1100	2560
55	F	UW	205	308	1025	2650	235	329	1151	2560
55	F	ER	213	533	1425	1830	214	535	1412	1800
56	F	IY	236	307	2670	3150	245	340	2700	3250
56	F	IH	231	417	2300	3000	239	410	2200	2910
56	F	EH	222	644	2250	3000	224	670	2300	2880
56	F	AE	224	784	1800	2750	234	820	1750	2890
56	F	AH	225	765	1300	2700	221	730	1390	2790
56	F	AA	225	834	1282	2800	212	850	1270	2760
56	F	AO	229	688	1029	2750	222	670	1040	2640
56	F	UH	251	427	1506	2640	240	460	1370	2610
56	F	UW	236	378	1416	2580	239	380	1430	2610
56	F	ER	230	460	1200	1909	225	410	1580	1800
57	F	IY	256	384	2860	3210	250	375	3000	3400
57	F	IH	230	460	2665	3140	233	467	2680	3150
57	F	EH	229	640	2400	2860	233	630	2530	3030
57	F	AE	233	700	2560	3150	225	675	2510	3145
57	F	AH	240	768	1440	2855	234	794	1447	2920
57	F	AA	227	978	1362	2724	233	933	1350	2610
57	F	AO	240	700	1080	2810	240	720	1090	2840
57	F	UH	243	500	1215	2870	239	500	1240	2860
57	F	UW	263	470	1000	2820	272	378	950	2990
57	F	ER	243	480	1410	1700	243	493	1580	1775
58	F	IY	268	320	2900	3200	263	290	2750	3050
58	F	IH	258	460	2380	2940	251	480	2260	2980

Table A.18: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
58	F	EH	246	640	2220	2900	250	670	2250	2960
58	F	AE	243	950	1970	2890	244	980	1950	2920
58	F	AH	251	750	1280	2760	258	770	1340	2800
58	F	AA	250	950	1130	3160	256	850	1150	2940
58	F	AO	242	530	870	2680	250	600	900	2770
58	F	UH	250	600	1225	2500	264	630	1320	2560
58	F	UW	258	440	1290	2530	269	460	1080	2640
58	F	ER	250	600	1500	2000	254	610	1520	1950
59	F	IY	234	280	2690	3040	261	280	2740	2980
59	F	IH	260	470	2500	3400	262	440	2480	3240
59	F	EH	242	730	2300	3100	260	750	2340	3120
59	F	AE	233	860	2070	2880	240	890	1920	2710
59	F	AH	257	770	1540	2840	257	800	1410	2860
59	F	AA	240	790	1250	3080	241	820	1210	2960
59	F	AO	234	408	695	3040	246	420	590	3100
59	F	UH	251	500	1230	2520	256	480	1230	2750
59	F	UW	263	419	1050	2850	278	390	1060	2800
59	F	ER	220	420	1720	1900	255	510	1680	1890
60	F	IY	208	270	2820	3450	225	250	2880	3350
60	F	IH	220	370	2530	3060	250	400	2600	3120
60	F	EH	214	640	2360	3020	219	650	2430	3040
60	F	AE	205	900	2090	3000	200	860	2160	2870
60	F	AH	214	750	1540	2800	214	770	1530	2780
60	F	AA	195	920	1350	2550	210	920	1470	2690
60	F	AO	194	720	1110	2420	200	700	1100	2780
60	F	UH	222	470	1200	2900	237	470	1190	2800
60	F	UW	240	380	980	3100	188	340	920	3050
60	F	ER	222	530	1670	2050	200	500	1720	1900
61	F	IY	258	310	2740	3200	262	262	2680	3170
61	F	IH	262	450	2310	3020	263	472	2270	2950
61	F	EH	245	640	1980	2920	235	700	2110	2940
61	F	AE	194	810	1860	2620	234	890	1800	2700
61	F	AH	230	710	1340	2780	245	740	1470	2940
61	F	AA	225	830	1020	2650	219	830	1095	2610
61	F	AO	240	600	850	2760	253	455	810	2750

Table A.19: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
61	F	UH	282	400	1070	2530	250	450	1050	2450
61	F	UW	260	290	670	2380	275	330	630	2460
61	F	ER	240	500	1630	2040	243	490	1580	2190
62	C	IY	228	460	3300	3950	200	400	3400	3850
62	C	IH	205	600	2550	4000	205	610	2500	4100
62	C	EH	225	600	2750	3600	210	760	2500	3850
62	C	AE	200	1000	2300	3900	200	800	2500	4050
62	C	AH	200	1000	1750	3550	223	1110	1690	4040
62	C	AA	205	1220	1560	3650	200	1300	1800	3450
62	C	AO	219	660	1100	3850	217	690	1090	3900
62	C	UH	206	620	1420	3700	220	620	1410	3520
62	C	UW	233	440	900	3900	200	400	650	3800
62	C	ER	210	610	2300	2900	200	450	2150	2550
63	C	IY	290	320	3500	4260	305	350	3400	4100
63	C	IH	322	640	3200	3660	325	650	3000	3800
63	C	EH	270	850	2900	3680	285	700	3120	3750
63	C	AE	256	1130	2560	3500	285	1140	2000	3560
63	C	AH	310	1130	1740	3670	300	1000	1800	3450
63	C	AA	265	1170	1500	3440	283	980	1300	3100
63	C	AO	265	530	1060	3450	272	540	1080	3000
63	C	UH	285	560	1440	3500	294	570	1450	3500
63	C	UW	333	350	1280	3650	290	340	1160	2950
63	C	ER	275	560	1740	2460	302	600	1800	2200
64	C	IY	240	380	3140	3700	258	310	3350	3650
64	C	IH	290	580	2760	3400	250	500	2660	3500
64	C	EH	250	780	2450	3400	240	672	2550	3400
64	C	AE	240	660	2900	3370	215	760	2850	3300
64	C	AH	250	880	1500	3200	243	850	1700	3250
64	C	AA	250	940	1380	2400	276	1200	1500	3160
64	C	AO	250	750	1250	3450	225	675	950	3240
64	C	UH	300	610	1500	3300	275	500	1370	3500
64	C	UW	256	300	1280	3150	250	400	1300	3700
64	C	ER	250	500	1540	1700	242	580	1620	1790
65	C	IY	291	410	3200	3800	264	420	3400	3900
65	C	IH	291	580	2900	3820	280	560	2840	3900

Table A.20: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
65	C	EH	292	810	2640	4200	270	780	2720	4100
65	C	AE	270	1080	2480	3950	245	1050	2420	4000
65	C	AH	286	970	1600	3950	250	800	1680	3800
65	C	AA	275	1040	1350	3850	250	1100	1460	4250
65	C	AO	286	770	1150	3950	273	710	1200	3900
65	C	UH	285	680	1420	3800	278	640	1350	3950
65	C	UW	300	420	1110	3640	280	505	1050	3400
65	C	ER	320	640	1940	2820	265	610	2100	2600
66	C	IY	330	460	2800	3550	333	490	2730	3550
66	C	IH	310	560	2500	3450	310	580	2500	3450
66	C	EH	286	800	2300	3750	310	835	2420	3740
66	C	AE	282	950	2150	3650	310	1000	2150	3700
66	C	AH	293	880	1700	3750	340	900	1600	3650
66	C	AA	299	990	1410	3750	280	1050	1320	3730
66	C	AO	285	770	940	3750	333	680	1020	3700
66	C	UH	322	550	1195	3750	350	550	1340	3500
66	C	UW	316	600	1200	3600	345	550	1100	3470
66	C	ER	310	805	1705	2420	310	710	1700	2400
67	C	IY	210	340	3400	4320	227	590	3610	4220
67	C	IH	235	680	3250	4380	220	440	3000	3790
67	C	EH	212	660	2900	3610	216	610	2760	3650
67	C	AE	214	1240	2700	3640	215	1050	2550	3550
67	C	AH	216	820	1470	3500	211	970	1410	3200
67	C	AA	218	1090	1380	3050	212	860	1250	2800
67	C	AO	211	800	1220	3700	214	640	1070	3000
67	C	UH	219	660	1360	3700	214	730	1500	3600
67	C	UW	220	620	1100	3250	216	600	1280	3650
67	C	ER	222	670	2130	2360	205	760	2240	2460
68	C	IY	253	330	3250	3720	262	340	3100	3400
68	C	IH	250	500	2500	3640	278	530	2630	3640
68	C	EH	255	710	2550	3560	250	750	2480	3470
68	C	AE	233	1140	2260	3640	245	1110	2230	3380
68	C	AH	256	770	1540	3500	257	800	1490	3300
68	C	AA	240	940	1400	3400	245	930	1370	3120
68	C	AO	240	530	860	3400	240	520	910	3420

Table A.21: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
68	C	UH	255	510	1250	3320	260	520	1140	3320
68	C	UW	274	360	660	3050	260	310	730	3500
68	C	ER	250	550	1500	1800	239	480	1650	1960
69	C	IY	250	300	2950	3600	270	320	3210	3600
69	C	IH	290	550	2610	3560	286	540	2570	3600
69	C	EH	280	700	2500	3580	263	600	2360	3400
69	C	AE	260	970	2400	3200	250	950	2270	3200
69	C	AH	270	780	1650	3350	250	720	1500	3240
69	C	AA	278	950	1200	2950	250	920	1080	2770
69	C	AO	262	790	1050	2900	250	750	1000	2500
69	C	UH	275	540	1430	3320	263	530	1580	3200
69	C	UW	295	420	1500	3010	260	450	1330	2840
69	C	ER	272	570	1880	2400	255	510	1610	1910
70	C	IY	235	280	2820	3400	244	317	3125	3500
70	C	IH	230	460	2520	3300	212	420	2480	3140
70	C	EH	235	657	2300	3300	232	672	2275	3300
70	C	AE	231	808	1950	3300	225	870	2000	3200
70	C	AH	236	706	1410	3200	211	720	1480	2880
70	C	AA	250	950	1350	3100	227	910	1360	2950
70	C	AO	203	700	1120	3070	230	690	920	2760
70	C	UH	250	475	1250	3150	212	460	1210	2750
70	C	UW	244	403	1100	2950	242	363	920	2900
70	C	ER	226	452	1580	1810	232	510	1550	1740
71	C	IY	230	280	3140	3830	250	300	3400	3950
71	C	IH	225	450	2700	3650	250	400	2840	3700
71	C	EH	215	580	2650	3550	220	620	2660	3770
71	C	AE	240	910	2370	3160	233	930	2350	3450
71	C	AH	250	770	1650	3420	230	690	1600	3350
71	C	AA	242	970	1450	3260	225	1010	1650	3150
71	C	AO	232	670	1160	3550	225	720	1260	3400
71	C	UH	216	500	1640	3580	250	450	1440	3500
71	C	UW	290	350	1160	3260	273	330	1090	3350
71	C	ER	240	430	1800	2400	233	470	1840	2400
72	C	IY	275	330	3050	3800	286	340	2860	3610
72	C	IH	280	500	2720	3360	230	600	2750	3550

Table A.22: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
72	C	EH	245	735	2450	3300	258	780	2560	3300
72	C	AE	235	940	2020	2580	232	1070	2320	2900
72	C	AH	268	860	1530	3100	256	970	1500	3050
72	C	AA	245	780	1250	3180	236	970	970	3120
72	C	AO	258	825	1210	3100	300	930	930	2900
72	C	UH	260	490	1460	2860	286	570	1320	2840
72	C	UW	275	470	1400	2800	286	370	1160	2800
72	C	ER	268	510	1660	2100	250	480	1700	1830
73	C	IY	295	380	3200	4000	267	350	3250	3700
73	C	IH	294	380	2960	3800	300	520	2900	3600
73	C	EH	280	670	2790	3600	275	620	2750	3500
73	C	AE	262	1070	2380	3100	275	1130	2320	3110
73	C	AH	290	700	1730	2960	270	725	1570	2900
73	C	AA	278	1110	1630	2780	280	1130	1400	3000
73	C	AO	292	580	930	2950	270	540	1070	3000
73	C	UH	300	450	1350	3000	320	520	1600	3150
73	C	UW	307	460	1460	3070	300	400	1700	3000
73	C	ER	300	540	1770	2040	286	540	2050	2300
74	C	IY	300	300	3250	3850	275	275	3280	3800
74	C	IH	286	570	2850	3400	267	485	2630	3450
74	C	EH	264	650	2880	3500	284	570	2900	3600
74	C	AE	260	1300	2280	3130	260	1300	2160	3300
74	C	AH	275	850	1540	3020	262	840	1580	2880
74	C	AA	250	1230	1300	3200	286	1090	1230	2980
74	C	UH	283	540	1420	3050	300	600	1440	2900
74	C	UW	280	390	1340	2830	284	340	1110	3080
74	C	ER	280	530	1650	1740	286	550	1660	1770
75	C	IY	265	370	2950	3400	290	370	2910	3480
75	C	IH	271	515	2740	3280	290	485	2600	3200
75	C	EH	262	630	2520	3150	272	565	2440	3120
75	C	AE	262	970	2030	2880	275	915	2130	2900
75	C	AH	270	810	1600	3230	280	760	1530	3180
75	C	AA	270	810	1350	2940	275	1000	1360	3000
75	C	AO	270	535	970	2960	275	550	1080	2850
75	C	UH	275	550	1420	3040	295	570	1500	3000

Table A.23: Vowel Data, M = Male, F = Female, C = Child

Talker		Vowel	Frequencies				Frequencies			
			F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
75	C	UW	283	510	1700	3020	278	500	1640	3050
75	C	ER	261	522	1830	2350	282	530	1800	2250
76	C	IY	320	350	3240	3760	344	344	3120	3640
76	C	IH	308	590	2760	3500	320	540	2900	3500
76	C	EH	307	830	2750	3650	308	800	2640	3540
76	C	AE	294	1140	2450	3230	239	1130	2550	3150
76	C	AH	310	930	1540	3120	315	950	1670	3150
76	C	AA	350	1190	1470	3150	314	1070	1460	2950
76	C	AO	300	910	1200	3180	330	830	1250	3250
76	C	UH	327	630	1310	3270	322	610	1550	3400
76	C	UW	345	520	1250	3460	334	500	1140	3380
76	C	ER	308	740	1850	2160	328	660	1830	2200

Bibliography

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[2] M. Pawlak and Y. F. Ng, "On kernel and radial basis function techniques for classification and function recovering," in *12th International Conference on Pattern Recognition*, vol. II, pp. 454–456, 1994.

[3] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Hemel Hemstead, Hertfordshire, England: Prentice-Hall International, 1982.

[4] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley, 1992.

[5] D. M. Monro, "Real discrete fast fourier transform. statistical algorithm as 97," *Applied Statistics*, vol. 25, pp. 166–172, 1976.

[6] K. Fukunaga and R. R. Hayes, "The reduced parzen classifier," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 423–425, April 1989.

[7] D. J. Hand, *Kernel Discriminant Analysis.* John Wiley & Sons Ltd., 1982.

[8] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," *Complex Systems*, vol. 2, pp. 321–355, 1988.

[9] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing unit," *Neural Computation*, vol. 1, pp. 281–293, 1989.

[10] F. Girosi and T. Poggio, "Networks and the best approximation property," *Biological Cybernetics*, vol. 63, pp. 169–176, 1990.

[11] M. J. D. Powell, "Radial basis functions for multivariable interpolation: A review," in *Algorithms for Approximation*, pp. 143–167, Oxford: Clarendon Press, 1987.

[12] R. Penrose, "A generalized inverse for matrices," *Proc. Cambridge Philos. Soc.*, vol. 51, pp. 406–413, 1955.

[13] T. Poggio and F. Girosi, "A theory of networks for approximation and learning," Tech. Rep. A.I. Memo No. 1140, C.B.I.P Paper No. 31, Massachusetts Institute of Technology, July 1989.

[14] D. Lowe, "Adaptive radial basis function nonlinearities and the problem of generalisation," in *First IEE International Conference on Neural Networks*, pp. 171–175, 1989.

[15] W. D. Beastall, "Recognition of radar signals by neural network," in *First IEE International Conference on Neural Networks*, (London, UK), pp. 139–142, 1989.

[16] J. Moody and C. Darken, "Learning with localized receptive fields," in *Proceedings of the 1988 Connectionist Models Summer School* (Touretzky, Hinton, and Sejnowski, eds.), Morgan-Kaufmann, Publishers, 1988.

[17] S. Chen, S. A. Billings, C. F. N. Cowan, and P. M. Grant, "Practical identification of NARMAX models using radial basis functions," *International Journal of Control*, vol. 52, pp. 1327–1350, December 1990.

[18] S. Chen, P. M. Grant, and C. F. N. Cowan, "Orthogonal least squares algorithm for training multi-output radial basis function networks," in *2nd International Conference on Artificial Neural Networks*, pp. 336–339, 1991.

[19] S.-R. Lay and J.-N. Hwang, "Robust construction of radial basis function networks for classification," in *1993 IEEE International Conference on Neural Networks*, vol. 3, pp. 1859–1864, 1993.

[20] M. W. Mak, W. G. Allen, and G. G. Sexton, "Comparing multi-layer perceptrons and radial basis functions," *Journal of Microcomputer Applications*, vol. 16, pp. 147–159, April 1993.

[21] M. T. Musavi, W. Ahmed, K. H. Chan, K. B. Faris, and D. M. Hummels, "On the training of radial basis function classifiers," *Neural Networks*, vol. 5, pp. 595–603, 1992.

[22] B. Lemarie, "Size reduction of a radial basis function network," in *Proceedings of the International Joint Conference on Neural Networks*, 1993.

[23] S. J. Bye, "Connectionist approach to sdh bandwidth management," in *3rd International Conference on Artifical Neural Networks*, pp. 286–290, 1993.

[24] V. Kadirkamanathan and M. Niranjan, "A function estimation approach to sequential learning with neural networks," *Neural Computation*, vol. 5, pp. 954–975, November 1993.

[25] R. Katayama, Y. Kajitani, K. Kuwata, and Y. Nishida, "Self generating radial basis function as neuro-fuzzy model and its application to nonlinear prediction of chaotic time series," in *Second IEEE International Conference on Fuzzy Systems*, pp. 407–414, 1993.

[26] S. Lee and R. M. Kil, "A gaussian potential function network with hierachically self-organizing learning," *Neural Networks*, vol. 4, pp. 207–224, 1991.

[27] J. Reynolds and L. Tarassenko, "Spoken letter recognition with neural networks," Tech. Rep. OUEL 1907/91, Robotics Research Group, University of Oxford, December 1991.

[28] P. K. Houselander and J. T. Taylor, "On the use of pre-defined regions to minimise the training and complexity of multi-layer neural networks," *First IEE International Conference on Artificial Neural Networks*, pp. 383–386, 1989.

[29] K. J. Cios, R. E. Tjia, N. Liu, and R. A. Langenderfer, "Study of continuous id3 and radial basis functoin algorithms for the recognition of glass defects," in *IJCNN-91-Seattle: International Joint Conference on Neural Networks*, vol. 1, pp. 49–54, 1991.

[30] Y. Lee, "Handwritten digit recognition using k nearest neighbor, radial-basis function and backpropagation neural networks," *Neural Computation*, vol. 3, pp. 440–449, Fall 1991.

[31] M. Carlin, "Radial basis function networks and nonlinear data modelling," in *Fifth International Conference. Neural Networks and Their Applications. Neuro Nimes 92*, pp. 623–633, 1992.

[32] J. Oglesby and J. S. Manson, "Radial basis function networks for speaker recognition," in *Proceedings of the 1991 International Conference on Acoustics, Speech and Signal Processing — ICASSP 91*, vol. 1, pp. 393–396, 1991.

[33] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[34] J. MacQueen, "Some methods for classificatoin and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics, and Probability* (L. M. LeCam and J. Neyman, eds.), (Berkeley), p. 281, U. California Press, 1967.

[35] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Monterey, California: Wadsworth Inc., 1984.

[36] R. Lippmann. Vowel data received by author through electronic mail, October 1993.

[37] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons Inc., 1990.

[38] T. Kohonen, *An Introduction to Neural Computing.* Berlin: Springer-Verlag, second ed., 1988.

[39] J. Ghosh, S. Chakravarthy, Y. Shin, C.-C. Chu, L. Deuser, S. Beck, R. Still, and J. Whitely, "Adaptive kernel classifiers for short-duration oceanic signals," in *IEEE Conference on Neural Networks for Ocean Engineering*, pp. 41–48, 1991.

[40] M. Vogt, "Combination of radial basis function neural networks with optimized learning vector quantization," in *1993 IEEE International Conference on Neural Networks*, pp. 1841–1846, 1993.

[41] S. Geva and J. Sitte, "Adaptive nearest neighbor pattern classification," *IEEE Transactions on Neural Networks*, vol. 2, pp. 318–322, March 1991.

[42] T. Poggio and F. Girosi, "Extensions of a theory of networks for approximation and learning: Dimensionality reduction and clustering," Tech. Rep. A.I. Memo No. 1167, C.B.I.P. Paper No. 44, Massachusetts Institute of Technology, April 1990.

[43] C. A. B. Smith, "Some examples of discrimination," *Ann. Eugen.*, vol. 13, pp. 272–282, 1947.

[44] P. Lachenbruch and M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 167–178, 1968.

[45] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 1979.

[46] A. K. Jain and M. D. Ramaswami, "Classifier design with parzen window," *Pattern Recognition and Artificial Intelligence*, pp. 211–228, 1988.

[47] J.-N. Hwang, S.-R. Lay, and A. Lippman, "Unsupervised learning for multivariate probability density estimation: Radial basis function and exploratory projection pursuit," in *1993 IEEE International Conference on Neural Networks*, vol. 3, pp. 1486–1491, 1993.

[48] J. Reynolds and L. Tarassenko, "Isolated word recognition with the radial basis function classifier," in *1991 International Conference on Neural Network*, pp. 345–349, 1991.