Kernel machines and classifier ensemble learning for biomedical applications

Peng, Li

2006

Peng, L. (2006). Kernel machines and classifier ensemble learning for biomedical applications. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/3452>

<https://doi.org/10.32657/10356/3452>

Nanyang Technological University

Downloaded on 26 Apr 2023 19:24:05 SGT

Kernel Machines and Classifier Ensemble Learning for Biomedical Applications

Peng Li

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2006

*To my parents and my wife,
who always give me love and support.*

Acknowledgments

I would like to express my deepest appreciation and sincere gratitude to my supervisors, Associate Professor Kap Luk Chan and Associate Professor Shankar M. Krishnan for their invaluable advice and great support in my study at Nanyang Technological University (NTU). It is them who encourage me to get through the tough time during the research. I not only learned specialized knowledge from them, but also the spirit of exploring scientific problems. This work would not have been possible without their excellent guidance and persistent encouragement.

I appreciate the School of Electrical and Electronic Engineering, NTU, Singapore, for kindly awarding me the Research Scholarship and providing excellent facilities for my research.

I would like to thank Dr. Lei Wang, Dr. Yan Sun, Dr. Sheng Fu, Dr. Marta Patricia Tjoa for their valuable suggestions and discussions. Also, I would like to give my many thanks to Ms. Juan Xu, Mr. Jia Yin Zhou, Ms. Wen Fang, Ms. Yan Gao, Ms. Miao Miao Zheng, Mr. Chee Khiang Ng and other research students and staff at BioMedical Engineering Research Center, NTU, Singapore. In addition, I am grateful to Dr. Choon Leong Tang from Singapore General Hospital for his clinical support and suggestions.

Finally, I would like to dedicate this thesis to my family and my wife, Ji Fang Jia, for their unlimited support and love.

Contents

Acknowledgments	i
Summary	vi
List of Figures	ix
List of Tables	xiii
List of Abbreviations	xv
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	3
1.3 Major Contribution of the Thesis	3
1.4 Organization of the Thesis	5
2 Support Vector Machines and the Imbalanced Data Problem	7
2.1 Introduction	7
2.2 Discriminative Support Vector Machines for Binary Classification . .	8
2.3 The Imbalanced Data Problem	15
2.3.1 Resampling	15
2.3.2 Using Different Costs to Two Classes	17
2.3.3 Recognition-based Approach	17
2.4 Recognition-based One-Class Support Vector Machines	18
2.4.1 Support Vector Data Description	20
2.4.2 ν -Support Vector Classifier	25
2.5 Experimental Results and Discussions	26

2.5.1	Evaluation Measure	27
2.5.2	The Influence of Class Imbalance to Discriminative <i>BSVCs</i>	29
2.5.3	The Performance of Recognition-based <i>OSVMs</i>	31
2.6	Summary of Findings	32
3	Hybrid Kernel Machine Ensemble	34
3.1	Introduction	34
3.2	Related Work	35
3.2.1	1.5-SVM	35
3.2.2	Multiple Classifier Ensemble	36
3.3	Hybrid Kernel Machine Ensemble	39
3.3.1	Hybrid Kernel Machine Ensemble	39
3.3.2	Binary <i>SVM</i> Training	41
3.3.3	One-class <i>SVM</i> Training	42
3.3.4	Fusion Rules for Integration of Hybrid <i>SVMs</i>	43
3.3.5	Estimating the Posterior Probability for Outputs of <i>SVMs</i>	46
3.4	Experimental Results and Discussions	47
3.4.1	Artificial Data	47
3.4.2	Real Data Sets	53
3.4.2.1	Breast Cancer Data Set	53
3.4.2.2	Pima Data Set	55
3.4.2.3	Glass Data Set	56
3.4.2.4	Blood Data Set	57
3.4.3	Comparison of <i>HKME</i> and 1.5 <i>SVM</i>	58
3.4.4	Summary	59
3.5	Concluding Remarks	60
4	Abnormal ECG Beat Annotation for Long-term Monitoring of Heart Patients	61
4.1	Introduction	61
4.2	Generalization and Imbalanced Data Problem in ECG Beat Annotation	65
4.2.1	Generalization Problem	65

4.2.2	Imbalanced Data Problem	66
4.3	Related Work	67
4.4	Concept Learning-based Approach	70
4.4.1	Proposed Methodology	70
4.4.1.1	Signal Preprocessing	71
4.4.1.2	Feature Extraction	74
4.4.1.3	Normalization	75
4.4.1.4	ν SVC-based Patient-adaptable ECG Beat Annotation	75
4.4.2	Experiments	75
4.4.2.1	Experimental Setting	76
4.4.2.2	Training Global Classifiers for Comparison	77
4.4.2.3	Training Local One-class Classifiers for Comparison .	78
4.4.2.4	Evaluation Measure	79
4.4.2.5	Results and Discussion	80
4.5	Hybrid Kernel Machine Ensemble Approach	86
4.5.1	Proposed Methodology	86
4.5.1.1	Signal Preprocessing, Feature Extraction and Normal- ization	87
4.5.1.2	HKME-based Patient-adaptable ECG Beat Annota- tion	88
4.5.2	Experiments	89
4.5.2.1	Experimental Setting	89
4.5.2.2	Results and Discussion	90
4.6	Concluding Remarks	97
5	HKME-based Abnormal Region Detection in Colonoscopic Images	99
5.1	Introduction	99
5.2	Related Work	102
5.2.1	Perceptual Image Segmentation	102
5.2.2	Colonoscopic Image Analysis	103
5.3	Proposed Methodology	105
5.3.1	Image Region Representation Using Multiple-size Patches . . .	105

5.3.2	Image Preprocessing	107
5.3.3	Feature Extraction	108
5.3.4	Learning <i>SVMs</i> for Image Patch Classification	109
5.3.4.1	<i>BSVC</i> Learning	109
5.3.4.2	ν <i>SVC</i> Learning	110
5.3.4.3	Decision Fusion Using <i>HKME</i>	110
5.4	Experimental Results and Discussions	111
5.4.1	Data Preparation	111
5.4.2	Evaluation Measure	111
5.4.3	Classification Based on Single Patch Size	112
5.4.4	Detecting Abnormal Region Using Learned <i>HKME</i>	112
5.4.4.1	ν <i>SVC</i> vs <i>BSVC</i> Using Single Patch Size	113
5.4.4.2	Multi-size Patch Ensemble of ν <i>SVCs</i> or <i>BSVCs</i>	113
5.4.4.3	<i>HKME</i> Using Single-size Patches	114
5.4.4.4	<i>HKME</i> Using Multi-size Patches	114
5.4.4.5	Observation from the Oracle	115
5.5	Concluding Remarks	116
6	Conclusions and Recommendations	119
6.1	Conclusions	119
6.2	Recommendations for Further Research	121
	Author's Publications	123
	Bibliography	125

Summary

Analyzing medical data for clinical diagnosis of abnormalities faces many challenges. Algorithmic processing and analysis of such data requires a computation model to map quantitative representations that are feasible to extract from the data into qualitative concepts. This research aims to explore the use of kernel machine and classifier ensemble to learn the qualitative concepts from tractable medical data. Through supervised learning from the data, diagnosis models can be extracted automatically and later applied to computer-assisted diagnosis.

From the point of pattern recognition, one challenge in computer-assisted diagnosis is a kind of imbalanced data problem in which the majority category is compactly clustered and the minority category is scattered in the input space. A discriminative model such as a binary Support Vector Machine (SVM) can be trained by manually balancing the data or compensating the imbalance using different costs to the two classes and it uses the information from both majority class and minority class. However, the performance of such a classifier model is usually poor due to the poorly represented minority class. A recognition-based approach such as a one-class SVM can do better than the discriminative approach in this case by modelling the well-represented majority class only because it avoids the problem caused by the inadequate representation of the minority class in binary SVMs. However, such a recognition-based model is not highly discriminative because the information from the minority class is left totally unused. An investigation into the two types of kernel machines reveal some nice properties of them. Exploiting the complementary natures of these two different types of models, the combination of them is expected to perform better than that of using either of them separately in the classification of this kind of imbalanced data set. Hence a new method is proposed to integrate these two kernel machines to form a hybrid kernel machine ensemble to address such imbalanced data problem aforementioned. Its good performance is justified on some toy data sets and two real medical applications, one is on Electrocardiogram (ECG) signal analysis, the other is on colonoscopic image analysis.

The first medical application is ECG signal analysis, which is very important for long-term monitoring of patients suffering from cardiovascular diseases. One challenge

in ECG beat analysis is large variation in the morphologies of ECG signals from different patients. The ranges of “normal” beats are different among the patients. Therefore, an ECG classifier finely tuned to the training data from a large group of patients may perform poorly when used to interpret the ECG beats from an individual patient. Furthermore, in the scenario of long-term monitoring of some heart patients, the normal ECG beats usually dominate the ECG recordings. Abnormal ECG beats appear differently in morphology and show large variation while normal ECG beats appear similar to each other and show less variation, which implies that the concept “normal” is more compact compared to that of the concept “abnormal” thus easier to be learned using fewer samples. A concept learning approach is hence proposed to solve the above problem in which a one-class SVM is trained using only normal ECG beats from a specific patient. Since the training data of the one-class SVM is more similar to the ECG beats of the specific patient than those beats from a large group of patients, the one-class SVM is expected to perform better than classifiers trained using the data from a large group of patients.

In the clinical examination of the ECG signals, the cardiovascular experts have to not only consider the special reference value from the patient (local information) but also the standard reference value from a large group of people (global information). So the standard reference value is also useful in the annotation of the ECG beats. Hence it is necessary to incorporate the global information into the concept learning model to further improve the annotation of the ECG beats. Hence the hybrid kernel machine ensemble is proposed to fuse these two different information. On one hand, a discriminative binary SVM is trained using ECG beats from different patients to embed the global information. On the other hand, a recognition-based one-class SVM is trained using only normal ECG beats from a specific patient to embed the local information. Integration of the two types of SVMs is expected to perform better than using either of them separately and help to improve the generalization. Experimental results using MIT/BIH arrhythmia ECG database show good performance of the proposed ensemble and support its feasibility in practical clinical application.

The aim of the second medical application is to detect abnormal regions in the colonoscopic images which is significant in the clinical screening for abnormalities through colonoscopy. The abnormal regions usually do not occupy the whole image

and vary in location, color, size and shape across images. A patch-based approach is adopted for abnormal region detection. But it is very difficult to determine the appropriate patch-size to use. Small patch size cannot capture sufficient information of the object and often lead to large detection error. Large patch size contains more information about the image regions that match its size and achieve more reliable detection, but fail to represent smaller regions. A new multi-size patch-based method is proposed to solve this problem in which multi-size patches are used simultaneously to represent the image region. At least some among all the patch sizes can better characterize the image region. Hence, an ensemble is constructed in which each classifier handles a patch size. The combination of classifiers trained with multiple-size patches can recognize the image region more effectively than using single-size patches only.

Many categories of abnormalities can be observed in colonoscopic images and they show large variations. Usually only a short interval of the colonoscopic image sequences shows abnormalities which implies the imbalanced data problem aforementioned. This can be solved by using the proposed hybrid kernel machine ensemble. A novel methodology of using hybrid kernel machine ensemble based on multi-size patches is hence proposed and applied successfully to colonoscopic image analysis.

List of Figures

2.1	Schematic illustration of the relation among expected risk, empirical risk and capacity term in inequality (2.4).	10
2.2	Kernel mapping and optimal separating hyperplane of SVM.	11
2.3	Typical decision boundaries of (a) two-class classifier and (b) one-class classifier in a 2-D toy problem.	18
2.4	Illustration of the kernel mapping of SVDD.	21
2.5	Illustration of the kernel mapping of νSVC	25
2.6	2×2 checkerboard data set.	27
2.7	The influence of class imbalance on the performance of $BSVC$ using 2×2 checkerboard data set (with negative samples as majority class).	30
2.8	The influence of class imbalance on the performance of $BSVC$ using 2×2 checkerboard data set (with positive samples as majority class).	30
2.9	The performance of νSVC in terms of BCR using 2×2 checkerboard data set, with negative samples or positive samples as target class for training respectively).	32
3.1	Approaches to designing an ensemble system [1].	36
3.2	The flowchart of hybrid kernel machine ensemble ($HKME$).	40
3.3	(A) Original target data set and (B) generated artificial outliers around the target class in a toy problem in 2-D space.	42
3.4	The comparison of different schemes using 2×2 checkerboard data set with different imbalance ratio.	48
3.5	The comparison of different fusion rules for $HKME$ using 2×2 checkerboard data set with different imbalance ratio.	50

3.6	The decision boundaries of νSVC , $BSVC$, and $HKME$ using 2×2 checkerboard data set (256 negative samples and 16 positive samples).	50
3.7	The decision boundary of $BSVC$ training using original data set and those of $BSVC$'s trained using (A) different costs to two classes, (B) SMOTE, (C) over-sampled data set and (D) under-sampled data set.	52
3.8	The performance of different schemes on Breast Cancer data sets. . .	54
3.9	The performance of different schemes on Pima data sets.	55
3.10	The performance of different schemes on Glass data sets.	56
3.11	The performance of different schemes on Blood data sets.	57
3.12	The performance of different schemes on 2×2 checkerboard data sets.	59
4.1	A typical normal ECG beat.	62
4.2	Examples of ECG beats using data from MIT/BIH arrhythmia database. A is a normal beat. The others are abnormal.	63
4.3	Scatterplot of ECG data of 8 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection.	67
4.4	Scatterplot of ECG data of all the ECG data from 44 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection.	68
4.5	Flowchart of the proposed framework for abnormal ECG beat annotation.	71
4.6	Scatterplot of ECG data of 4 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection. "TN" are the first 5 minutes of each recording while "TT" are the last 15 minutes of each recording).	72
4.7	Scatterplot of ECG data of 4 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection. "TN" are the first 5 minutes of each recording while "TT" are the last 15 minutes of each recording.	73
4.8	Comparison of the annotation results of the νSVC and $RSVC$ in each recording of the test sets in terms of BCR.	81

4.9	The relation between the training set size of νSVC model and its performance in terms of BCR (Mean $\pm$ Standard Deviation). From 1 to 10, the length of the ECG signal used in the training set decreases from 15 (15/1) minutes to 1.5 (15/10) minutes.	82
4.10	Scatterplot of ECG data of 2 test sets in MIT/BIH arrhythmia database showing the first two principal components of PCA projection.	83
4.11	Flowchart of the proposed framework for abnormal ECG beat annotation.	87
4.12	Comparison of the annotation results of the global $RSVC$, local νSVC and $HKME$ (with DET fusion rule) in each recording of the test sets in terms of BCR.	91
4.13	The influence of training set size to the performance of the $HKME$ (DET) in terms of BCR.	93
4.14	The relation between the improved BCR of the $HKMEs$ ($DETs$) over the global $RSVC$ (A) and local $\nu SVCs$ (B) and the plain disagreement measure.	94
4.15	The relation between the improved BCR of the $HKMEs$ ($DETs$) over the global $RSVC$ (A) and local $\nu SVCs$ (B) and the Double Fault measure (DFM).	95
4.16	The relation between the improved BCR of the $HKMEs$ ($DETs$) over the global $RSVC$ (A) and local $\nu SVCs$ (B) and the correlation coefficient measure (COR).	95
5.1	Colonoscope and colonoscopy system	100
5.2	A normal colonoscopic image and 5 colonoscopic images with different types of abnormalities.	101
5.3	The flowchart of $HKME$ -based approach for abnormal region detection in colonoscopic images.	105
5.4	Patch-based image region representation.	106
5.5	Two examples of patches from colonoscopic images. The component L, a and b of the image patches are illustrated listed in a row.	108
5.6	Detection results of 4 colonoscopic images. The regions in white are normal regions detected and the regions in black are abnormal ones.	116

5.7	Detection results of 4 colonoscopic images. The regions in white are normal regions detected and the regions in black are abnormal ones.	117
-----	--	-----

List of Tables

2.1	Decision table of two-class classification outcome for calculating the evaluating measure.	27
3.1	Original, training and test data set used in this study.	53
3.2	BCR (average $\pm$ standard deviation in percentage) of each scheme using Breast Cancer data set.	54
3.3	BCR (average $\pm$ standard deviation in percentage) of each scheme using Pima data set.	55
3.4	BCR (average $\pm$ standard deviation in percentage) of each scheme using Glass data set.	57
3.5	BCR (average $\pm$ standard deviation in percentage) of each scheme using Blood data set.	58
3.6	BCR (average $\pm$ standard deviation in percentage) of each approach using 2×2 checkerboard data set.	59
4.1	Description of the ECG training and Test data sets for reference. * Note that only “Normal” ECG beats are used in training the local νSVC s	77
4.2	Full classification matrix for calculating the evaluation criteria used in this study.	79
4.3	Results (average $\pm$ standard deviation) of abnormal ECG beat annotation (in percentage).	81
4.4	Annotation results of the νSVC and $RSVC$ in each recording of the test sets.	82

4.5	Comparative results of testing time and number of Support Vectors (average $\pm$ standard deviation) of abnormal ECG beat annotation. The test time is the total time of annotation of the 18 recordings (24010 beats).	85
4.6	Results (average $\pm$ standard deviation) of abnormal ECG beat annotation (all in percentage).	90
4.7	Number and percentage of test sets in which <i>HKME</i> outperforms both global <i>RSVC</i> and local ν <i>SVC</i> among all 22 test sets.	90
4.8	Annotation results of the <i>RSVC</i> , ν <i>SVC</i> and <i>HKME</i> with DET fusion rule in each recording of the 22 test sets.	92
5.1	Results of image patch classification.	112
5.2	Results of abnormal region detection using single patch sizes.	113
5.3	Detection results (in terms of <i>BCR</i>) of abnormal region detection using different patch sizes and ensemble schemes. The ensembles are constructed using same types of <i>SVMs</i> , <i>BSVC</i> or ν <i>SVC</i>	114
5.4	Detection results (in terms of <i>BCR</i>) of abnormal region detection using same patch sizes by <i>HKME</i>	114
5.5	Detection results (in terms of <i>BCR</i>) of abnormal region detection using different patch sizes and <i>HKME</i> . Row <i>A + *</i> are the results of ensembles using all 3-size patches learned by <i>BSVC</i> s plus 1-size or 2-size patches learned by ν <i>SVC</i> (s) (1 for 48×48 , 2 for 32×32 , 3 for 16×16). Row <i>ALL</i> are the ensemble results using all 3-size patches and both <i>BSVC</i> and ν <i>SVC</i>	115

List of Abbreviations

AAMI	Association for the Advancement of Medical Instrumentation
ACR	Average Classification Rate
AVG	Average
BCR	Balanced Classification Rate
BSV	Bounded Support Vector
BSVC	Binary Support Vector Classifier
CFD	Coincident Failure Measure
COR	Correlation Coefficient Measure
D	Dimensional
DET	Decision Template
DFM	Double Fault Measure
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
FN	False Negatives
FP	False Positives
GM	Geometric Mean
HKME	Hybrid Kernel Machine Ensemble
KKT	Karush-Kuhn-Tucker condition
LDC	Linear Discriminant Classifier
LSVC	Linear Binary Support Vector Classifier
MAX	Maximum
MFNN	Multi-layer Feed-forward Neural Network
MIN	Minimum
MOG	Mixture of Gaussians
NMC	Nearest Mean Linear Classifier
NN	Neural Network
NNDD	Nearest Neighbor Data Description

ORA	Oracle
OSVM	One-class Support Vector Machine
PCA	Principal Component Analysis
PDM	Plain Disagreement Measure
PROD	Product
QDC	Quadratic Discriminant Classifier
RBF	Radial Basis Function
RSVC	Binary Support Vector Classifier with RBF kernel
SEN	Sensitivity
SMOTE	Synthetic Minority Oversampling Technique
SPE	Specificity
SRM	Structural Risk Minimization
SVDD	Support Vector Data Description
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
USV	Unbounded Support Vector
VC	Vapnik Chervonenkis theory
ν SVC	One-class Support Vector Classifier

Chapter 1

Introduction

1.1 Motivation

Computer-assisted diagnosis means assistance to the physicians for solving a diagnostic problem with computers and is an active area of research. The technologies behind it include pattern recognition, signal processing, computer vision and machine learning, etc. Analyzing medical data for clinical diagnosis of abnormalities faces many challenges. Algorithmic processing and analysis of such data requires a computation model to map quantitative representations that are feasible to extract from the data into qualitative concepts. One can mimic the diagnosis process of medical experts through a quantitative computational model constructed from the qualitative concepts of the medical experts. However, the modelling of the diagnosis process is very difficult because the expertise of the medical experts is often fuzzy and hard to quantify. Instead, this research aims to explore the use of soft computing techniques to learn the qualitative concepts from tractable medical data to alleviate the problems faced by medical experts in quantitatively defining the basis of their judgements and in the search for quantitative models. Through supervised learning from the data, diagnosis models can be constructed automatically and later applied to computer-assisted diagnosis.

One challenge in computer-assisted diagnosis is a kind of imbalanced data problem in which the majority class is compactly clustered and the minority class is scattered in the input space. For example, in patient monitoring or medical screening for

abnormalities using images or signals, the signal morphologies from normal activities are usually similar and the data from this class can be easily collected (majority class), while those from abnormal activities may exhibit various morphologies and are more difficult to be collected (minority class). This is a two-class classification task in which a classifier has to be inferred from a set of annotated training data to discriminate the data in the abnormal class from those of the normal one.

Two-class classification is usually solved by training a discriminative model such as Binary Support Vector Classifiers (*BSVCs*). *BSVC* is a typical Support Vector Machine (*SVM*) which is originally developed for two-class classification. Designed using structural risk minimization principle, *BSVCs* have shown better performance than traditional classifiers [2]. Using imbalanced data sets, the discriminative *BSVC* can be trained by manually balancing the data or compensating the imbalance using different costs to the two classes and it uses the information from both majority class and minority class. However, its performance is usually poor due to the poorly represented minority class.

One approach to deal with the imbalanced data problem is the recognition-based model instead of the traditional discriminative classifiers [3]. The learning problem is treated as a one-class classification task rather than a two-class classification problem. The one-class classification task can be solved using kernel-based one-class Support Vector Machines (*OSVMs*) such as a one-class Support Vector Classifier (ν *SVC*) [4] or Support Vector Data Description (*SVDD*) [5]. These one-class classifiers are trained using only the data from one class (usually the majority class) [6], and therefore avoid the problem faced by a discriminative model caused due to the inadequate representation of the minority class. The one-class classifiers are expected to do better than the discriminative approach on imbalanced data sets by modelling the well-represented majority class only. However, recognition-based one-class classifiers are not highly discriminative because the information from the minority class is left totally unused in the training of these classifiers. Hence, the one-class classifiers seldom outperform the traditional two-class classifiers in practical applications when the data from two classes are available [7, 8].

On one hand, two-class classifier *BSVC* benefits from the information from two classes while suffering from insufficient representation of the minority class in an

imbalanced data set. On the other hand, one-class classifier *OSVM* benefits from the more precise representation of the majority class while it is not highly discriminative. This motivates the development of a new types of learning algorithm which is in between the one-class classifier and two-class classifier benefitting from the advantages of both classifiers so that it can perform better than using either of the existing two types of classifiers on this kind of imbalanced data sets. Such learning algorithm is expected to improve the classification performance in many medical applications, such as medical screening for abnormalities or patient monitoring.

1.2 Objectives

The major aim of this study is to develop effective algorithms to solve a kind of imbalanced data problem which exists in many medical applications, such as detecting abnormal regions in colonoscopic images and abnormal Electrocardiogram (*ECG*) beat annotation for long-term monitoring of heart patients.

The following aspects of medical data analysis are discussed and studied in this work:

- Hybrid Kernel Machine Ensemble (*HKME*) algorithm for a kind of imbalanced data problem where one category of data is compactly clustered and the other category of data is scattered in the input space.
- Concept learning-based and *HKME*-based abnormal ECG beat annotation approaches for long-term monitoring of heart patients.
- *BSVC* ensemble-based and *HKME*-based abnormal region detection approaches using multi-size patches in colonoscopic images.

1.3 Major Contribution of the Thesis

In this thesis, a novel learning algorithm, i.e. *HKME* is developed to address a kind of imbalanced data problem in many medical applications where one category of data is compactly clustered and the other category of data is scattered in the input space. The proposed *HKME* algorithm is analyzed and evaluated using both

artificial and real data sets. In addition, the *HKME* algorithm is applied to two real medical applications, one is abnormal ECG beat annotation for long-term monitoring of heart patients and the other is abnormal region detection in colonoscopic images.

The major contributions of this work are summarized as follows:

- A new learning algorithm, *HKME* is proposed to address a kind of imbalanced data problem where the majority class is compactly clustered and the minority class is scattered in the input space. *HKME* is an ensemble of two types of *SVMs*, i.e. a discriminative *BSVC* and a recognition-based *OSVM*. Exploiting the complementary properties of these two different types of kernel machines, combining of them using an ensemble achieves better performance than that of using either of them alone. The proposed *HKME* can be regarded as a learning algorithm in between the one-class classifier and two-class classifier, which can be called an *one-and-half classifier*. The properties of *HKME* are evaluated using some artificial data sets and real data sets. Experimental results show its feasibility for the intended task.
- A concept learning-based approach is proposed for patient-adaptable abnormal ECG beat annotation for long-term monitoring of heart patients. This is an attempt to address the difference between a patient group and an individual patient. The generalization of the ECG beat classifier trained using the data from a large pool of patients is usually poor when used to analyze ECG beats for different patient. A concept learning model such as *OSVM* can be trained using several minutes of normal ECG beats from each patient on long-term monitoring to learn the concept “normal”. The learned model can then be used to annotate the abnormal ECG beats in the ECG recordings of the same patient. The proposed method is good in generalization, easy to implement and fast to calculate.
- An *HKME*-based approach is developed for patient-adaptable abnormal ECG beat annotation for long-term monitoring of heart patients. This is motivated by the fact that a physician considers both the standard reference value of the patient group and the specific reference value of the patient in clinical monitoring of heart patients. Such *HKME* consists of a *OSVM* trained using some

normal ECG beats from a patient and a *BSVC* trained using the data from a large group of patients. The ensemble of such two *SVM*s can perform better than both the global *BSVC* and the local *OSVM*.

- A new method is proposed to detect abnormal regions in colonoscopic images using multi-size patch-based *BSVC* ensemble. Multi-size patches provides multiple-level representation of image content. At least some among all the patch size can better characterize the abnormal regions. A *BSVC* classifier is trained for each patch size. The final decision can be made by fusing the decision made by each classifier. Preliminary experimental results show that the proposed multi-size patch-based *BSVC* ensemble is able to produce more perceptually agreeable detection in colonoscopic images.
- Building on the success of using multi-size patch representation, *HKME* is also applied to the abnormal region detection in colonoscopic images based on the multi-size patch-based image representation and in order to solve the imbalanced data problem in the abnormal region detection. Experimental results show the good performance of the proposed *HKME* algorithm using multi-size patches and suggest its potential in real clinical screening for abnormalities in colonoscopic images.

1.4 Organization of the Thesis

The thesis is organized as follows:

- In Chapter 2, a review of two-class discriminative *SVM* and one-class recognition-based *SVM* is presented. They are applied to a type of imbalanced data sets where one category of data is compactly clustered and the other category of data is scattered in the input space. Their performances on such imbalanced data sets are evaluated using some artificial data sets.
- In Chapter 3, the *HKME* is introduced by combining a discriminative *BSVC* and a recognition-based *OSVM*. Some related work, including 1.5*SVM* and

multiple classifier ensemble, is reviewed. The properties of *HKME* are investigated and evaluated using a checkerboard data set and four UCI data sets.

- Chapter 4 describes two approaches for abnormal ECG beat detection for long-term monitoring of heart patients, concept-learning based approach and *HKME* based approach. Their performance are evaluated using the MIT/BIH arrhythmia database [9].
- Chapter 5 presents the application of *HKME* in colonoscopic image analysis. A multi-size patch-based representation is firstly developed for abnormal region detection. A *BSVC* ensemble is proposed to fuse the detecting results using different patch sizes and it is then extended to *HKME*-based framework.
- Finally, conclusions are drawn and recommendation for future research is summarized in Chapter 6.

Chapter 2

Support Vector Machines and the Imbalanced Data Problem

2.1 Introduction

Kernel machines refer to the algorithms that map the data from input space into higher-dimensional feature space where simple linear methods are used to process the mapped data. The mapping is usually nonlinear and is implemented implicitly through the kernel trick. The nonlinear mapping increases flexibility of the algorithm and the kernel trick simplifies the computation efficiently. Due to these advantages, many kernel methods have been developed in domain of machine learning, such as Support Vector Machines (*SVMs*) [2, 10–12], Kernel-based Principal Component Analysis (KPCA) [10, 13], Kernel-based Linear Discriminant Analysis (KLDA) [14], Kernel-based Independent Component Analysis (KICA) [15] and Kernel-based Nearest Neighbor Classifier [16], etc.

One of the most widely used kernel machines, *SVM*, was originally developed for two-class classification. Based on the principle of structural risk minimization, discriminative binary *SVMs* (*BSVC*) usually perform well in real applications [17–20]. However, *SVMs* also suffer from some fundamental problems in statistical pattern recognition, such as imbalanced data problem [21], in which the number of the training data from one class is significantly larger than that of the other class in a two-class classification task.

One possible solution to the imbalanced data problem is to use recognition-based approach instead of the generally used discriminative two class classification [3]. Recognition-based approach is based on one-class classification in which only the data from one class (usually the majority class) are used to train a classifier [6], which is significantly different from the traditional two-class classifiers learned using the data from both classes. There are two kernel machines that can be used as one-class classifiers, including one-class Support Vector Classifier (ν SVC) [4] and Support Vector Data Description (SVDD) [5]. These one-class SVMs (OSVC) are trained using the data from the majority class only. Therefore, they avoid the problem of inadequate representation of the minority class which is usually the cause of imbalanced data problem. However, the performance of one-class classifiers are seldom superior to the traditional two-class classifiers in real applications [7]. One reason might be that the data distribution of majority and minority classes is not suitable for modelling as a one-class classification problem. Another reason may be due to the fact that only the data from one-class are used to infer the classifier in one-class classification and no information about the other class is used. Hence the one-class classifiers are constructed to describe the one-class training data rather than for discrimination purpose.

In this chapter, the fundamentals of both discriminative two-class SVMs and recognition-based one-class SVMs are introduced. Their performance on a particular type of imbalanced data problem is evaluated using some artificial data sets.

2.2 Discriminative Support Vector Machines for Binary Classification

SVMs are probably the most widely used kernel methods, which have been increasingly used in many biomedical applications, such as [17–20]. This approach is motivated by statistical learning theory [22] and it can be applied to classification, regression and concept learning. SVMs are originally developed for two-class classification (or binary classification) and it can be extended for multiple-classification as well [10].

The task of two-class classification is to infer a classifier from a training set. The two classes are labelled as +1 and -1 respectively. The training set is a set of labelled

training samples or patterns

$$X = \{x_i \in R^d | i = 1, 2, \dots, N\} \quad (2.1)$$

where N is the number of samples in the training set. Each sample x_i is represented by a feature vector of d dimensions and labelled as $y_i \in \{+1, -1\}$. The classifier can be represented by a function $f(x) : x \rightarrow y$. A label y can be obtained for each pattern x by the classifier. It is assumed that the training data and test data are drawn from the same distribution $P(x, y)$. The optimal function f can be found by minimizing the expected risk

$$R(f) = \int r(f(x), y) dP(x, y) \quad (2.2)$$

where r is a loss function. For example, $r(f(x), y) = |f(x) - y|$ whose value is 0 for correct classification ($f(x) = y$) and 1 for incorrect classification. Such a loss is called 0/1 loss.

In practice, the underlying probability distribution $P(x, y)$ is usually unknown. Therefore, the risk R cannot be minimized directly. One can approximate the risk by minimizing the empirical risk

$$R_{em}(f) = \frac{1}{N} \sum_{i=1}^N r(f(x_i), y_i) \quad (2.3)$$

The empirical risk R_{em} will converge to the expected risk R when the number of training samples tends to infinity ($N \rightarrow \infty$). However, overfitting may occur when the number of training samples are small [23]. Instead, one can approach the expected risk while avoids overfitting using the Vapnik Chervonenkis (VC) theory and the structural risk minimization (SRM) principle [2, 22].

Let $F_1 \subset F_2 \subset \dots \subset F_m$ denote a family of function classes with non-decreasing VC dimension and $f_1, f_2, \dots, f_m$ be the solutions of the empirical risk minimization (2.3) in the function class F_i . Let v be the VC-dimension of the function class F . For any $v < N$ and with probability of at least $1 - \eta$, the expected risk is bounded by the following inequality [2, 22]

$$R(f) \leq R_{em}(f) + R_{ca}(v, N, \eta) \quad (2.4)$$

Figure 2.1: Schematic illustration of the relation among expected risk, empirical risk and capacity term in inequality (2.4).

where the first item in the right hand side is the empirical risk in (2.3) and the second item R_{ca} is called the capacity term

$$R_{ca} = \sqrt{\frac{v(\ln \frac{2N}{v} + 1) - \ln \frac{\eta}{4}}{N}} \quad (2.5)$$

The relation among the expected risk, the empirical risk and the capacity term in inequality (2.4) is illustrated in Figure 2.1. The goal is to minimize the expected risk, $R(f)$, which consists of two items, the empirical risk R_{em} and the capacity term R_{ca} . Obviously, one cannot minimize the expected risk by only minimizing the empirical risk. Instead, one has to minimize the empirical risk and the capacity term at the same time. Their tradeoff can be controlled by the VC-dimension v since both are related to v .

In practice, the bound on the expected risk is often difficult to compute. Fortunately, the decision functions in SVMs are restricted to hyperplanes whose VC-dimension can be bounded in terms of another quantity, called the “margin” [2].

Given the two-class training set $X = \{x_i \in R^d | i = 1, 2, \dots, N\}$ with N samples which may not be separable, the data are mapped to another feature space using a mapping M in which the mapped data $M(x)$ can be separated by an optimal

separating hyperplane expressed as

$$f(x) = (w \cdot M(x)) - b \quad (2.6)$$

in which w is the weight vector, b is a bias item. $(\cdot)$ is a inner product. Such mapping is illustrated in Figure 2.2.

Figure 2.2: Kernel mapping and optimal separating hyperplane of SVM.

The “margin” is defined as the minimal distance of a sample to the decision hyperplane $f(x)$. w and b can be scaled so that the closest point to the hyperplane satisfy $|w \cdot M(x) - b| = 1$. Then the margin can be calculated using two samples from opposite classes $M(x_1)$ and $M(x_2)$ which have $w \cdot M(x_1) - b = 1$ and $w \cdot M(x_2) - b = -1$ respectively,

$$\frac{w}{\|w\|} \cdot (M(x_1) - M(x_2)) = \frac{2}{\|w\|} \quad (2.7)$$

It has been proven that minimizing the “margin” can minimize the capacity term in (2.4) [10]. This can be formulated as an optimization problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (2.8)$$

with constraints

$$y_i((w \cdot M(x_i)) - b) \geq 1, i = 1, 2, \dots, N \quad (2.9)$$

For noisy data, one can introduce some slack variables θ_i to relax the constraints in (2.9):

$$y_i((w \cdot M(x_i)) - b) \geq 1 - \theta_i, \quad \theta_i \geq 0, \quad i = 1, 2, \dots, N \quad (2.10)$$

The optimization problem in (2.8) can be reformulated as

$$\min_{w, b, \theta} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \theta_i \right\} \quad (2.11)$$

where $C > 0$ is a regularization parameter to control the tradeoff of the empirical error and the capacity term.

The minimization problem in (2.11) is called primal in optimization theory. Its first item is related to the capacity item R_{ca} in (2.4) and the second item is the empirical risk R_{em} . Therefore minimizing (2.11) can minimize the expected risk R . This problem can be solved by introducing Lagrange Multipliers $\beta_i \geq 0$ and $\gamma_i \geq 0$, $i = 1, 2, \dots, N$, for the constraints in (2.10). The following Lagrangian can be formulated

$$L(w, b, \theta, \beta, \gamma) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \theta_i + \sum_{i=1}^N \beta_i (1 - \theta_i - y_i((w \cdot M(x_i)) - b)) - \sum_{i=1}^N \gamma_i \theta_i \quad (2.12)$$

The goal is to minimize (2.12) with respect to primal variables w and b and to maximize it with respect to the Lagrange multipliers β_i and γ_i . At the optimal point, the partial derivatives of L to w , b and θ_i are equal to 0 respectively. The following equations can therefore be derived:

$$w = \sum_{i=1}^N y_i \beta_i M(x_i) \quad (2.13)$$

$$\sum_{i=1}^N y_i \beta_i = 0 \quad (2.14)$$

$$C - \beta_i - \gamma_i = 0 \quad (2.15)$$

Substituting (2.13) (2.14) and (2.15) into (2.12), one can get the dual problem

$$\max_{\beta} \left\{ \sum_{i=1}^N \beta_i - \frac{1}{2} \sum_{i,j=1}^N y_i y_j \beta_i \beta_j (M(x_i) \cdot M(x_j)) \right\} \quad (2.16)$$

with constraints

$$\sum_{i=1}^N y_i \beta_i = 0 \quad (2.17)$$

and

$$C \geq \beta_i \geq 0, i = 1, 2, \dots, N \quad (2.18)$$

This is a quadratic programming problem, which can be solved using some standard algorithms, such as Sequential Minimization Optimization [24]. Related codes can be found in [25].

At the optimal point, the Karush-Kuhn-Tucker (KKT) complementarity conditions [26] have to be satisfied, i.e. the items in (2.12) related to Lagrange multipliers β_i and γ_i approaches zero, so that the Lagrangian (2.12) is the same as the cost function in the primal problem (2.11). Hence the following equations can be derived

$$\beta_i (1 - \theta_i - y_i ((w \cdot M(x_i)) - b)) = 0 \quad (2.19)$$

$$\gamma_i \theta_i = 0 \quad (2.20)$$

Analysis of the optimal solution of (2.16) is quite informative. Three cases for each β_i are resulted due to the KKT optimality conditions:

- $0 < \beta_i < C$. Then $0 < \gamma_i < C$ holds due to (2.15). From (2.20), $\theta_i = 0$ holds. Equation (2.19) leads to $y_i ((w \cdot M(x_i)) - b) = 1$. This means that a point x_i with $0 < \beta_i < C$ locates on the hyperplane. Such a point is called an Unbounded Support Vector (USV).
- $\beta_i = C$. From (2.19), $y_i ((w \cdot M(x_i)) + b) = 1 - \theta_i$. Under these constraints, a point with $\beta_i = C$ locates in the margin. Such a training pattern is referred to as a Bounded Support Vector (BSV) for it has hit the upper bound.

- $\beta_i = 0$. Then $\gamma_i = C$ due to (2.15). From (2.20), $\theta_i = 0$ holds, which means the corresponding point is correctly classified.

It can be seen that only the data with $\beta_i = 0$ contribute to the decision function in (2.6). They are the support vectors (including both BSVs and USVs) which are either on the margin or inside the margin. Therefore the solution is sparse.

Furthermore, it can be found that only inner product is calculated in (2.16) and (2.6). No explicit mapping M is needed. Such an inner product can be replaced using a kernel function

$$K(x_i, x_j) = M(x_i) \cdot M(x_j) \quad (2.21)$$

provided that this kernel $K(x_i, x_j)$ satisfies the Mercer's theorem. Then (2.16) and (2.6) can be reformulated using (2.13) as follows

$$\max_{\beta} \left\{ \sum_{i=1}^N \beta_i - \frac{1}{2} \sum_{i,j=1}^N y_i y_j \beta_i \beta_j K(x_i, x_j) \right\} \quad (2.22)$$

$$f(x) = \sum_{i=1}^N y_i \beta_i K(x_i, x) - b \quad (2.23)$$

Among all possible kernels, Gaussian kernel is one of the most widely used

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}} \quad (2.24)$$

The bias item can be calculated by using the fact that for all USVs with $0 < \beta_i < C$, the slack variable $\theta_i = 0$. For any such USV x_s , the bias item can be calculated using (2.19), (2.13) and (2.21) as:

$$b = \sum_{i=1}^N y_i \beta_i K(x_i, x_s) - y_i \quad (2.25)$$

Designed using the structural risk minimization (SRM) principle, SVM has shown good performance in many applications. However, SVM also suffer from some fundamental problems in statistical pattern recognition tasks, such as imbalanced data problem which is introduced in the next section.

2.3 The Imbalanced Data Problem

The imbalanced data problem has received considerable attention in recent year from the machine learning community. This problem occurs when the size of the training set from one class is significantly larger than that of the other class in a two-class classification setting. This problem is often encountered in real applications such as medical screening for abnormalities, image retrieval, oil spill in satellite images [27, 28] etc. In applications such as medical screening for abnormalities, the data of the normal class can be easily obtained. On the other hand, the data of the abnormal class are more difficult to be collected compared to those of the normal ones. Therefore, the data from the abnormal class (usually the minority class) cannot represent well to its true distribution compared to the other class (usually the majority class, in here this is the normal class). It can be assumed that the minority class is positive and the majority class is negative in this chapter and it is formulated as a simple binary classification problem. Hence it is logical to use discriminative binary classifiers such as binary SVMs. However, such classifiers are designed to minimize the overall misclassification rate on the training set, their performances degrade when encountering such a highly imbalanced data set.

Some attempts have been reported to deal with the imbalanced data problem, which can be classified into the following 3 approaches [8, 21].

2.3.1 Resampling

The first approach is resampling the training data set to make it balanced, e.g. [29, 30]. Resampling is probably the most extensively studied approach, which consists of two main techniques:

1. Undersampling: The data from majority class are down-sampled so that the size of the majority class matches the size of the minority class. One can randomly select the samples from the majority class to balance the training data set [21]. But the resulted balanced data set may not faithfully represent the distribution of the original data set. There are also some attempts to improve the undersampling technique. For example, Kubat *et al* proposed a one-sided

selection method to remove noisy or unreliable examples from the majority class [30].

The size of the balanced data set obtained using undersampling is less than that of the original data set, which reduces the computation complexity of the training process. But the problem is that some of the information may be lost if down-sampling is not done properly.

2. Oversampling: The data from minority class are over-sampled so that the size of minority class matches the size of the majority class. Similar to random undersampling, random oversampling has shown effective in improving the classification [8, 21]. Though duplicating the samples of minority class has some detrimental effect on classification accuracy, it is proven that the duplicate rates has to be very high to be truly harmful [31].

There are also some attempts to improve the performance of oversampling. For example, Nickerson *et al* proposed to address the between-class imbalance problem using guided oversampling by taking the within-class imbalances into consideration [32]. Chawla *et al* developed a Synthetic Minority Oversampling Technique (SMOTE) by generating artificial data (the nearest neighbors of the original minority data) [33].

It is unclear which of these two is more effective in solving the imbalanced data problem [29, 34]. Therefore, some attempts have also been made to combine these two approaches [29, 33, 35].

It is shown that original class distribution is usually not the best for learning [36]. In Weiss and Provost's experiments, the balanced data set generally leads to results which are no worse than or superior to those of using natural class distribution [36], although it does not always produce the optimal results. This observation is also supported by [37]. The distribution of training data set is changed by resampling. So whether this is beneficial to the classification depends on the data distribution.

2.3.2 Using Different Costs to Two Classes

The second category is to compensate for the class imbalance by altering the costs of the minority and majority classes in the training of classifiers. For example, Karakoulas *et al* proposed an algorithm called ThetaBoost, which is a boosting algorithm with unequal loss functions, to deal with imbalanced data set [38]. Leskovec *et al* extended linear programming boosting to handle imbalanced data problem by optimizing a cost criterion which can be adapted to reflect the imbalance of the data set.

Some attempts have also been made to compensate the class imbalance by using different costs to the two classes in the training of SVMs [39,40]. Raskutti *et al* used different penalizing factors for two classes and resampling for SVM in [7]. Wu *et al* proposed the class-boundary alignment algorithm to deal with imbalanced data problem for SVM [41].

2.3.3 Recognition-based Approach

The third category is to use recognition-based approach instead of discrimination-based learning by leaving one of the two classes totally unused (usually the minority class). The recognition-based approach resembles that of a density estimation without finding the true density explicitly. This is an extreme case where only the data from one class is used to construct the learning model. For example, Japkowicz proposed to use an autoencoder to solve the imbalanced data problem [3]. This approach works well when the majority class can be well modelled by a novelty detector such as an autoencoder. However, the recognition-based approach is usually outperformed by discrimination-based approach as a consequence of excluding the information from the minority class in the training of the model [7]. There are also some other recognition-based classifiers, such as one-class support vector machines which is introduced in the next section.

2.4 Recognition-based One-Class Support Vector Machines

One-class classification is also known as novelty detection, outlier detection and concept learning [6]. The problems in one-class classification is different from those in conventional two-class classification. In one-class classification it is assumed that only information of one of the classes, **the target class**, is available. The training data is from only one-class and there is no information about the other class, **the outlier class** available. The task of one-class classification is to define a boundary around the target class, such that it accepts as much of the targets as possible and excluding the outliers as much as possible (Figure 2.3-(b)).

Figure 2.3: Typical decision boundaries of (a) two-class classifier and (b) one-class classifier in a 2-D toy problem.

The philosophy behind concept learning is in agreement with the way that human-beings learn a concept. Suppose one expects to teach a child the concept of “car”. One only needs to give him or her some examples of cars and not necessary to give the examples of non-“car”, such as truck, bus or train. That is to say, people can learn a concept using only the examples of the target class. Of course, the information about non-target or outliers is helpful to improve the discrimination of the target class from the non-target one. However, using the examples from only the target class is sufficient to learn the concept of the target and recognize whether a new pattern

belongs to the concept of “target class”.

Since the one-class classifiers are trained using the data from one class only, a threshold is usually set so that the decision boundary of the classifier can enclose the target samples as much as possible. This is usually difficult without the information from the other class. One way is to reject some target to form a tighter boundary as shown in Figure 2.3 (b). Then the threshold can be determined by the errors over the target class only [6].

One-class classification has been used in many fields. For example, Murshed *et al* used a fuzzy ARTMAP-based classification system for detecting cancerous cells [42]. Hojjatoleslami *et al* employed a RBF network for density estimation in the detection of microcalcifications in mammograms [43]. Hayton *et al* applied ν -Support Vector Classifier to novelty detection of jet engine vibration spectra [44]. Manevitz and Yousef used One-Class Support Vector Machine for document classification [45]. Tax *et al* employed support vector data description in pump failure detection [5] and image retrieval [46]. Xin *et al* employed support vector data description for speaker recognition [47, 48].

There are several one-class classifiers. A survey of the one-class classifiers can be found in [49, 50]. They are based on three main approaches. Namely, density estimation, boundary methods and reconstruction methods [6]:

1. Density estimation: It directly estimates the density of the target class to form a model to represent the data. The generally used models include Parzen, Gaussian and mixture of Gaussian etc. For example, in Parzen density estimation method, the density for the targets are estimated based on Parzen-window, and the test point is classified by the maximum posterior [51]. Generally speaking, this approach works well when the sample size is sufficiently high and a flexible model is used. However, when the model does not fit the data very well, a large bias may be resulted.
2. Reconstruction: A model is constructed to fit to the data. Models such as K-means, Learning Vector Quantization, Principal Component Analysis and neural networks etc. can be used to model the data in this approach. For examples, k-means is one of the well-known clustering algorithms. It tries to

model the training set with k prototypes measured by the squared Euclidean distance. This approach may fail when the model does not fit the data well.

3. Boundary approach: It focuses on the boundary of the data. Given a set of data, only a closed boundary around the target set is optimized. K-centers, nearest neighbor, ν -Support Vector Machine and Support Vector Data Description etc. belong to this approach. Compared to density estimation methods, it avoids the estimation of the complete probability density. Therefore, it may be possible to learn from data when the exact target density distribution is unknown or when only limited number of samples are available. So it is more efficient than density estimation methods [6].

Typical decision boundaries of two-class classifier and one-class classifier are illustrated in Figure 2.3. The decision boundary of two-class classifier is supported by the samples of both classes and it utilizes the information from both classes. The two-class classifier is trained for “discrimination” purpose. But the decision boundary of one-class classifier is formed using only the data from one class. The one-class classifier is trained to “recognize” the target samples rather than for “discrimination” purpose. Therefore, classification performance of one-class classifiers is usually worse than two-class classifiers when the data from both classes are available [7].

There are two well-known kernel methods for one-class classification currently available. One is called ν -Support Vector Classifier (ν SVC) [4]. Another is called Support Vector Data Description (SVDD) [5]. It has been proven that these two methods are equivalent to each other when Gaussian RBF kernel is used [4].

2.4.1 Support Vector Data Description

SVDD is a kind of support vector machine [5] which can be used as a one-class classifier. Given a set of target data with n samples $X = \{x_i \in R^d | i = 1, 2, \dots, n\}$. It may be difficult to find a decision boundary directly in the original d -dimensional input space. Therefore, a nonlinear mapping M is sought to map X into some high dimensional kernel-induced feature space, in which a hypersphere is sought to enclose the mapped target data $M(X)$ with a smallest radius R and center $\mathbf{c}$. Figure 2.4

Figure 2.4: Illustration of the kernel mapping of SVDD.

illustrates the nonlinear kernel mapping. The problem becomes

$$\min_{\{R, c, S\}} \left\{ R^2 + \frac{1}{\nu N} \sum_{i=1}^N S_i \right\} \quad (2.26)$$

subject to

$$\| M(x_i) - c \|^2 \leq R^2 + S_i, \quad i = 1, 2, \dots, N \quad (2.27)$$

where S_i ($S_i \geq 0$) are some slack variables to allow soft boundaries – some target data are allowed to lie outside of the hypersphere, thus to control the tradeoff between two types of errors. $\nu \in (0, 1]$ is a regularization parameter used to control the tradeoff between the size of the hypersphere and the errors. In fact, it is the upper bound of the fraction of target data located outside the hypersphere which will be shown later.

The above problem can be solved by constructing a Lagrangian. Introducing constraints (2.27) to cost function (2.26), the Lagrangian is formulated as

$$\begin{aligned} L(R, c, S, \gamma, \beta) = & R^2 + \frac{1}{\nu N} \sum_{i=1}^N S_i - \sum_{i=1}^N \gamma_i S_i \\ & - \sum_{i=1}^N \beta_i (R^2 + S_i - \| M(x_i) - c \|^2) \end{aligned} \quad (2.28)$$

where $\beta_i \geq 0$ and $\gamma_i \geq 0$ are Lagrange multipliers. To solve this primal problem, L has to be minimized with respect to R , c and S , while maximized with respect to γ and β .

Setting the partial derivatives of L with respect to the primal variables R , c and S_i to zero, yields

$$\sum_{i=1}^N \beta_i = 1 \quad (2.29)$$

$$c = \sum_{i=1}^N \beta_i M(x_i) \quad (2.30)$$

$$\frac{1}{\nu N} - \beta_i - \gamma_i = 0, \quad i = 1, 2, \dots, N \quad (2.31)$$

Since $\gamma_i \geq 0$ and $\beta_i \geq 0$, constraint (2.31) can be rewritten as

$$0 \leq \beta_i \leq \frac{1}{\nu N}, \quad i = 1, 2, \dots, N \quad (2.32)$$

At the optimal point, the Karush-Kuhn-Tucker (KKT) complementarity conditions [26] have to be satisfied, i.e. the items in (2.28) related to Lagrange multipliers γ_i and β_i approaches zero, so that the Lagrangian (2.28) is the same as the cost function in the primal problem (2.26). Hence we have

$$\beta_i (R^2 + S_i - \|M(x_i) - c\|^2) = 0 \quad (2.33)$$

$$\gamma_i S_i = 0 \quad (2.34)$$

Substituting (2.29), (2.30) and (2.31) into (2.28) leads to the so-called dual problem in optimization theory, which can be regarded as an equivalent problem to the primal problem.

$$\max_{\beta} \left\{ \sum_{i=1}^N \beta_i (M(x_i) \cdot M(x_i)) - \sum_{i,j=1}^N \beta_i \beta_j (M(x_i) \cdot M(x_j)) \right\} \quad (2.35)$$

with constraints

$$\sum_{i=1}^N \beta_i = 1 \quad (2.36)$$

$$0 \leq \beta_i \leq \frac{1}{\nu N}, \quad i = 1, 2, \dots, N \quad (2.37)$$

Define function $K(\cdot, \cdot)$ as

$$K(x_i, x_j) = M(x_i) \cdot M(x_j) \quad (2.38)$$

Then equation (2.35) becomes

$$\min_{\beta} \left\{ \sum_{i,j=1}^N \beta_i \beta_j K(x_i, x_j) - \sum_{i=1}^N \beta_i K(x_i, x_i) \right\} \quad (2.39)$$

with same constraints as (2.35). The cost function of dual problem (2.39) is convex and quadratic in terms of the unknown parameters β_i . This problem can be solved by quadratic programming, for which some standard algorithms such as sequential minimization optimization can be employed [4, 6].

Analysis of the optimal solution of (2.39) is also quite informative. Similar to two class case, three cases for each β_i are resulted due to the KKT optimality conditions:

- $0 < \beta_i < \frac{1}{\nu N}$. Then $0 < \gamma_i < \frac{1}{\nu N}$ holds due to (2.31). From (2.34), we have $S_i = 0$. Equation (2.33) leads to $\|M(x_i) - c\|^2 = R^2$. This means that a point x_i with $0 < \beta_i < \frac{1}{\nu N}$ locates on the hypersphere, then it is an Unbounded Support Vector (USV).
- $\beta_i = \frac{1}{\nu N}$. From (2.33), we have $\|M(x_i) - c\|^2 = R^2 + S_i$. Under these constraints, a point with $\beta_i = \frac{1}{\nu N}$ locates outside of the hypersphere. This is also a Bounded Support Vector (BSV).
- $\beta_i = 0$. Then $\gamma_i = \frac{1}{\nu N}$ due to (2.31). From (2.34), we have $S_i = 0$. From (2.33), we have $\|M(x_i) - c\|^2 < R^2$, which means the corresponding point lies inside the hypersphere.

Therefore, USVs lie on the decision boundary, BSVs lies outside of the boundary – they are regarded as outliers, and all the other training patterns lies inside the decision boundary.

Obviously, the center of the sphere is the linear combination of data patterns (2.30). Note that there are only a few training patterns that can satisfy the equality of constraints (2.27), which are those that on or outside the boundary of the hypersphere

– the USVs and BSVs, whose coefficient β_i are non-zero. Only these patterns are needed to describe the hypersphere.

Through quadratic programming, the Lagrangian (2.39) is optimized with respect to β . The center of the hypersphere c and multiplier γ_i can be calculated by (2.30) and (2.31) respectively using the optimal solution β , . Because $\|M(x_i) - c\|^2 = R^2$ holds for all the USVs, the radius R can be calculated by choosing any of the USVs x_s

$$R = [K(x_s, x_s) + \sum_{i,j=1}^N \beta_i \beta_j K(x_i, x_j) - 2 \sum_{i=1}^N \beta_i K(x_i, x_s)]^{-2} \quad (2.40)$$

Given a new pattern z , the decision function is

$$\begin{aligned} f(z) = R^2 - \|M(z) - c\|^2 = R^2 - K(z, z) \\ - \sum_{i,j=1}^N \beta_i \beta_j K(x_i, x_j) + 2 \sum_{i=1}^N \beta_i K(z, x_i) \end{aligned} \quad (2.41)$$

If the value of the decision function is greater than zero, the new sample lies inside the hypersphere. Hence it is classified as a target. Otherwise it is classified as an outlier.

Similar to binary SVMs, kernel function can be used in SVDD. Although nonlinear mapping has been used to improve the effectiveness of the hyperspherical description, neither does the explicit nonlinear mapping $M(\cdot)$ appear in the dual problem of SVDD (2.39), nor in the decision function (2.41). They are expressed completely in terms of $K(x_i, x_j)$, which is the advantage of kernel method. In fact, since the problem is stated completely in terms of the inner products of the vectors, the inner products of the patterns can be replaced by a kernel function (2.38), provided that this kernel $K(x_i, x_j)$ satisfies Mercer's theorem [6]. Among all possible kernels, Gaussian kernel is one of the choices

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}} \quad (2.42)$$

It provides a very flexible description, which has been proven by Tax [5]. Because Gaussian kernel only depends on $x_i - x_j$, $K(x, x)$ is constant 1. Therefore (2.39) becomes

$$\min_{\beta} \sum_{i,j=1}^N \beta_i \beta_j K(x_i, x_j) \quad (2.43)$$

subject to the same constraints as (2.39). The decision function (2.41) can be reformulated as follows using (2.40)

$$f_d(z) = \sum_{i=1}^N \beta_i [K(x_i, z) - k(x_i, x_s)] = \sum_{i=1}^N \beta_i K(x_i, z) - b \quad (2.44)$$

where the bias term b is

$$b = \sum_{i=1}^N \beta_i K(x_i, x_s) = \sum_{i=1}^N \beta_i e^{-\frac{\|x_i - x_s\|^2}{\sigma^2}} \quad (2.45)$$

Here the SVDD decision function behaves as a template-matching detector in the mapped feature space. Since $\beta_i \neq 0$ holds only for those USVs and BSVs, these patterns forms a known template. Given a new pattern, it is compared with only the USVs and BSVs in the mapped feature space. Those patterns that are similar with all of the USVs and BSVs tend to have large negative value in (2.44) and they are more likely to be outliers. While those patterns different from all of the USVs and the BSVs tend to have large positive value in (2.44) and they are more likely to be targets.

2.4.2 ν -Support Vector Classifier

Figure 2.5: Illustration of the kernel mapping of ν SVC.

Another algorithm for estimating the support of a data distribution in kernel feature space is ν SVC. The kernel mapping is different from that of SVDD. The

target data are mapped into a higher-dimensional space called feature space $M(x)$ in which the dot product can be computed using some kernel function. The mapped target data are separated from the origin corresponding to the outliers with maximum margin using a hyperplane as shown in Figure 2.5, which can be found by solving the following problem

$$\min_{w, S_i, b} \frac{\|w\|^2}{2} + \frac{1}{\nu N} \sum_{i=1}^N S_i - b \quad (2.46)$$

subject to

$$w \cdot M(x_i) - b + S_i \geq 0, \quad \gamma_i \geq 0, \quad i = 1, 2, \dots, N \quad (2.47)$$

where S_i is a slack variable, $\nu \in (0, 1]$ is a regularization parameter to control the effect of outliers and allows for target samples falling outside the decision boundary. The decision function corresponding to the hyperplane is

$$f(x) = w \cdot M(x) - b \quad (2.48)$$

By similar analysis as in SVDD, this problem can be solved as a quadratic programming problem:

$$\min_{\beta} \sum_{i,j=1}^N \beta_i \beta_j K(x_i, x_j) \quad (2.49)$$

subject to

$$\sum_{i=1}^N \beta_i = 1 \quad (2.50)$$

$$0 \leq \beta_i \leq \frac{1}{\nu N}, \quad i = 1, 2, \dots, N \quad (2.51)$$

which is exactly the same as the dual problem (2.43) in SVDD when Gaussian kernel is used. Since Gaussian kernel is used in this study, νSVC and $SVDD$ are equivalent to each other.

2.5 Experimental Results and Discussions

In order to investigate the properties of the discriminative $BSVC$ and recognition-based $OSVM$ on the imbalanced data sets, the following experiments were conducted using a checkerboard data set similar to the one in [41]. The checkerboard data are

Table 2.1: Decision table of two-class classification outcome for calculating the evaluating measure.

Ground Truth	Classification Outcome	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	Truth Negative

within a unit square area in the two-dimensional space as shown in Figure 2.6. The negative samples (majority class) occupy the two diagonal squares of the checkerboard and the positive samples (minority) occupy in a 2×2 square around the negative samples. The data are uniformly distributed. The distribution of the checkerboard data is in agreement to the assumption that the majority class is compactly clustered and the minority class is scattered in the input space.

Figure 2.6: 2×2 checkerboard data set.

2.5.1 Evaluation Measure

The decision table of two-class classification outcome for calculating the evaluating criteria used in this study is illustrated in Table 2.1. Four cases can be observed from the classification results, i.e. true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Let A_+ and A_- denote the classification accuracy rate

of positive class and negative class respectively.

$$A_+ = \frac{TP}{TP + FN} \quad (2.52)$$

$$A_- = \frac{TN}{TN + FP} \quad (2.53)$$

The most commonly used measure is the Average Classification Rate (ACR) which is the fraction of all correctly classified samples among all the samples, regardless of positive or negative.

$$ACR = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.54)$$

In imbalanced data sets, the negative (majority) class dominates data set. Generally used ACR is not valid in evaluating the performance of the classifiers in such an imbalanced data set. For example, if a classifier is trained to classify all the data as negative beats, it has $A_- = 100\%$ and $A_+ = 0\%$, but the ACR is still high due to the small number of positive samples.

Since ACR is not suitable to the imbalanced data sets, another measure called Balanced Classification Rate (BCR) is used in this study.

- BCR is the algebraic mean of A_+ and A_- .

$$BCR = \frac{A_+ + A_-}{2} \quad (2.55)$$

This measure has been used in evaluating the performance of classifiers in imbalanced data sets [40, 52]. This measure is more suitable for evaluating the performance of the classifiers than the generally used ACR . Only when both A_+ and A_- have large value can BCR have a large value. Therefore, the use of BCR can have a balanced assessment of the classifiers in this kind of imbalanced data sets as the BCR favors both lower false positives and false negatives.

This measure is similar to the Geometric Mean (GM) defined by Kubat [30], which also favors both lower false positives and false negatives. The GM is defined as the geometric mean of A_+ and A_- .

$$GM = \sqrt{A_+ \cdot A_-} \quad (2.56)$$

There are also some other possible measures, such as the Receiver Operating Curves [53] etc.

2.5.2 The Influence of Class Imbalance to Discriminative *BSVC*s

In order to show the influence of imbalanced data set to the performance of discriminative *BSVC*s, the following experiments were conducted.

In the first experiment, the size of negative training data in a 2×2 checkerboard data was fixed as 128, the size of the positive training data was reduced from 128 to 8, with corresponding imbalance ratio (majority to minority) increased from 1 : 1 to 32 : 1). The number of test data consists of 1000 positive data and 1000 negative data. *BSVC*s with RBF kernel were trained using these data. The hyper-parameters of the *BSVC*s were optimized using 3-fold cross validation on the training set. The experiment was repeated 10 times and the average value and standard deviation of the *BCRs* achieved by *BSVC*s are reported in Figure 2.7.

It is observed that *BSVC* performs well when the training data set is balanced which is expected. But its performance deteriorates gradually as the imbalance ratio increases. It indicates that discriminative *BSVC* suffers from class imbalance. When the number of minority samples is very small, the data from this class cannot represent its true distribution well. This can be observed from the larger variation in performance of the *BSVC* when the imbalance ratio is large. Hence the poor performance of *BSVC* in highly imbalanced data sets is resulted from the inadequate representation of the minority class.

It is unclear whether *BVC*s also suffer from the imbalanced data problem when the two classes are adequately represented while the samples from two classes are imbalanced. Therefore, the size of negative training data was still fixed as 128 in the second experiment, while the size of the positive training data was increased from 128 to 2048, with corresponding imbalance ratio (minority to majority) increased from 1 : 1 to 1 : 16). The other settings are the same as the first experiment. The experimental result is illustrated in Figure 2.8.

It is observed that the result is similar to that in the first experiment. It shows that the discriminative *BSVC* also suffer from the class imbalance when the data are

Figure 2.7: The influence of class imbalance on the performance of *BSVC* using 2×2 checkerboard data set (with negative samples as majority class).

Figure 2.8: The influence of class imbalance on the performance of *BSVC* using 2×2 checkerboard data set (with positive samples as majority class).

more precisely represented while the two classes are highly imbalanced.

In summary, it is shown that the discriminative *BSVC* suffer from the class imbalance problem, hence some measures have to be taken to solve this problem.

2.5.3 The Performance of Recognition-based $OSVM$ s

It is mentioned in section 2.3 that one approach to address the class imbalance is to use recognition-based model instead of discriminative model by training a one-class classifier using the data from the majority class only. However, one-class classifiers seldom outperform two-class classifiers when the data from two class are available. One reason is that the one-class classifier is designed for describing the majority class rather than for discrimination purpose, leaving the information from another class totally unused. Another reason may be that the concept to be learned is not suitable for description by the one-class classifiers. For example, in patient monitoring, the concept “normal” signal is suitable for description by a one-class classifier while the concept “abnormal” signal is not. Because the “normal” class is usually compactly clustered in the input space, while the “abnormal” class is usually scattered in the input space. Furthermore, there is no clear boundary between “normal” and “abnormal” class. If the one-class classifier is to enclose the scattered “abnormal” data, it will also include some “normal” data. But it may be possible to construct a one-class classifier to enclose the compactly clustered “normal” data.

The negative class of the checkerboard data in Figure 2.6 seems more suitable to be described by a one-class classifier than the positive class since it is compactly clustered. This can be proved in the following experiment.

In the experiment, the negative samples and positive samples in a 2×2 checkerboard data set were taken as target class respectively to train a νSVC . The number of training data was varied from 8 to 1024. The number of test data consists of 1000 positive data and 1000 negative data. The hyper-parameters of the νSVC were optimized using artificially generated data set described in section 3.3.3. The experiment was repeated 10 times and the average value and standard deviation of the $BCRs$ achieved by νSVC s are reported in Figure 2.9.

It can be observed that the νSVC trained using compactly clustered negative samples outperforms that of using scattered positive samples. This supports the claim that compactly clustered negative class is more suitable to train one-class classifiers than scattered positive class.

Furthermore, the performance of νSVC is directly related to number of training

Figure 2.9: The performance of ν SVC in terms of BCR using 2×2 checkerboard data set, with negative samples or positive samples as target class for training respectively).

samples. It seems that $128 \sim 256$ negative samples have been quite good to train a ν SVC in this data set, whose performance is even better than the ν SVC trained using 1024 positive samples. It has been pointed out that more data is needed in one-class classification than a two-class classification [6]. So the number of training samples of ν SVC should be large enough to have a good description of the target class. In imbalanced data sets, a compactly clustered majority class will be suitable for the one-class classifiers to learn.

2.6 Summary of Findings

In this chapter, the fundamentals of discriminative $BSVCs$ and recognition-based $OSVMs$ are reviewed. Though discriminative $BSVCs$ were reported to perform well in many applications, they still suffer from a fundamental problem – the imbalanced data problem. Recognition-based approach is one of the techniques that are designed to deal with imbalanced data sets, which is trained using the data from one class only. This approach can be implemented using one-class $SVMs$ such as $SVDD$ or ν SVC. But it has been reported that recognition-based approach seldom outperform discriminative $BSVCs$ in real applications. This is due to the majority data may not

be suitable to be formulated as one-class problem.

Sets of data with different imbalance ratios are generated based on a checkerboard distribution shown in Figure 2.6. These checkerboard data sets are used to evaluate the performance of discriminative *BSVCs* and the recognition-based *OSVMs* when the training data set is imbalanced. Experimental result shows that the performance of discriminative *BSVCs* deteriorates as the imbalance ratio increases. Furthermore, it is shown that the *OSVM* performs well when the compactly clustered majority class is used to train the one-class model. But the performance of *OSVM* is inferior when the scattered class is modelled by the *OSVM*. It is expected that *OSVM* will show good performance when a compactly clustered majority class is modelled in an imbalanced data set.

Chapter 3

Hybrid Kernel Machine Ensemble

3.1 Introduction

It has been shown in the previous chapter that discriminative two-class *SVM* and recognition-based one-class *SVM* have problems in dealing with imbalanced data sets. Note that a two-class classifier *BSVC* benefits from the information from two classes while suffering from inadequate representation of the minority class. One-class classifier *OSVC* benefits from the more precise representation of the majority class but is not highly discriminative. There is a need to develop a classifier which is in between the one-class classifier and two-class classifier. Such a classifier can be named as ***one and half classifier***. By exploiting the different properties of the two types of kernel machines, an ensemble is constructed to combine these two types of kernel machines. Such an ensemble is called Hybrid Kernel Machine Ensemble (*HKME*). *HKME* is designed to benefit from both discriminative *BSVC* and recognition-based *OSVC* and such an ensemble is expected to perform better in some applications such as in imbalanced data sets or in other cases where there is a need to combine the information from these two types of kernel machines.

In this chapter, *HKME* is presented and applied to a kind of imbalanced data problem, in which the majority class is compactly clustered and the minority class is scattered in the input space. Such a situation is often encountered in real-life applications. For example, in patient monitoring, the signal morphologies from normal activities are similar and the data from this class can be easily collected (majority

class), while those from abnormal activities may exhibit various morphologies and are more difficult to be collected (minority class). In this case, the majority class satisfies the assumption of one-class classification where the majority class is well-represented and compactly clustered, it avoids the problem in *BSVC* caused by inadequate representation of the minority class. Furthermore, a discriminative *BSVC* can be trained by manually balancing the data so that it can utilize the information from both majority class and minority class. Exploiting the complementary nature of these two different types of models, the combination of them using *HKME* is expected to perform better than that of using either of them separately in the classification of this kind of imbalanced data set. Experimental results using some artificial data sets and real data sets show the feasibility of the proposed *HKME*.

3.2 Related Work

3.2.1 1.5-SVM

Although *SVDD* and ν *SVC* have been used in many applications, there are few applications in which one-class SVM outperforms binary SVMs when data from two classes are available [7]. This is due to the fact that only the data from one-class is used in one-class SVMs. In order to improve the performance of one-class SVMs, some attempts have been made to incorporate the information of the other class (outliers) into one-class SVMs. Such algorithms are in between the one-class classifier and two-class classifier, hence can be named as 1.5 classifier. For example, Tax proposed to improve the data description using negative examples for *SVDD* when the negative examples are available [6]. The cost function in (2.26) is changed to

$$\min_{\{R, c, S\}} \left\{ R^2 + \frac{1}{\nu_+ N_+} \sum_{i=1}^{N_+} S_i + \frac{1}{\nu_- N_-} \sum_{j=1}^{N_-} S_j \right\} \quad (3.1)$$

where ν_+ and ν_- are regularization parameters for two classes, N_+ and N_- are the number of examples in two classes. It has shown to reject the outliers better, but it does not fit tightly around the rest of the targets any more [6]. Similar idea has also been introduced to ν *SVC*. For example, an algorithm called Modified Support

Vector Novelty Detector was proposed in [54] and another algorithms called 1.5SVM was developed for image retrieval [55].

These algorithms seems promising, however few applications have been reported. This may be due to the difficulties of model selection. Further investigation is necessary.

3.2.2 Multiple Classifier Ensemble

Multiple classifier ensemble is also known as mixture of experts, classifier fusion and combination of multiple classifiers, etc [56]. It is a mechanism to combine a set of classifiers so that the resultant ensemble has superior classification performance to that of using each individual classifier alone.

A multiples classifier system belongs to one of the two categories, classifier fusion and classifier selection [57]. The individual classifiers are trained in parallel and their outputs are aggregated to make the final decision in a classifier fusion. While in classifier selection, the final decision is made by selecting the output of a single classifier with the largest confidence from a group of classifiers.

Figure 3.1: Approaches to designing an ensemble system [1].

The combination of the classifiers can be made using different training data sets, different feature subsets, various types of individual classifiers and fusion rules, as illustrated in Figure 3.1 [1].

1. Ensemble based on different training data sets: The ensemble using different training data sets is also called data level fusion, which is capable of producing an ensemble of diverse classifiers [58]. Different training data sets can be

produced by random partitioning the original data sets into several exclusive subsets, resampling the original data set to design bootstrap subsets –bagging, gradually building more “difficult sample” subsets from the original data set –boosting [59] etc. For example, a support vector machine ensemble using bagging and boosting was investigated in [60]

2. Ensemble based on different feature subsets: It is called feature level fusion [61,62]. This scheme shows effectiveness especially when the dimensionality of the feature vector is high compared to the number of the data samples. For example, an ensemble based on random subsets of features called attribute bagging was proposed in [63]. Feature subset level ensemble construction based on Genetic Algorithms was investigated in [64].
3. Ensemble of different types of classifiers or same types of classifiers trained using different strategies. It is also shown to be useful in many applications. For example, multi-layer neural networks can be trained using different initial weights, different structure, different training time or different parameters etc [57,65,66].
4. Ensemble of classifiers using different fusion rules: It is also called decision level fusion. The ensemble rules depend on the form of output of the classifiers. A survey on the combination methods for different output levels can be found in [67]. Generally speaking, the outputs of the classifiers can be categorized into hard level and soft level [68].

(a) Hard level: It means that the classification results (labels) given by the classifiers are used in the ensemble. It includes Majority voting [65,69] and weighted voting [60] etc. Though simple to calculate, these methods have shown quite effective in many applications.

(b) Soft level: It means that the confidence of the classifiers is used in the ensemble. The confidence is usually the posterior probability of the classifiers.

Some of the ensembles need no training after the base classifiers in the ensemble have been trained. For example, Maximum, Minimum, Median,

Averaging (or Sum), Product [46, 70, 71], Fuzzy integral [72, 73], mixture of experts [74].

The other ensembles need to be trained. For examples, Decision Template [56] and stacking [75] etc.

It is unclear whether training of ensemble is more helpful to the classification performance [76]. An experimental comparison between fixed and trained fusion rules for multimodal personal identity verification is reported in [77]. It is shown that the advantage of trained rules depends on the quality and the size of the training set.

It was shown that combining classifiers trained on different feature sets is very useful and best performance is achieved by combining both different feature sets and different classifiers in [71]. A survey on the design of multiple classifier systems can be found in [78].

Some attempts have been made to explain why multiple classifier ensemble outperforms individual classifiers. One explanation is based on the bias-variance decomposition of the error [51]. It indicates that ensembles not only reduce variance [79] but also bias [80]. A bias-variance Analysis of Support Vector Machines Ensemble can be found in [81]. Another explanation is under the framework of large margin classifiers. It shows that an ensemble is able to enlarge the margins, thus improve the generalization [82].

Although the ensembles have been used successfully in many applications, there is still no solid theoretical foundation available on how to design the optimal ensemble till now. There have been some attempts to link the performance of the ensemble to data complexity [83], independence of the classifiers [84] and diversity [69, 85, 86]. Independence between individual classifiers is usually regarded useful in classifier fusion. But it is reported that a negatively-correlation of the individual classifiers in an ensemble is preferable for improving the classification using majority vote rule when combining dependent classifiers [84].

One important condition for the success of an ensemble is that the outputs of individual classifiers to the same inputs must be diverse [85, 86]. The diverse individual classifiers can be obtained by using different training data sets, different feature

subsets, various types of individual classifiers etc. as illustrated in Figure 3.1.

Some attempts have been made to exploit the relationship between diversity and ensemble performance. For example, diversity measures were used to evaluate the performance of an ensemble of linear classifiers using Boosting and Bagging in [87]. It is indicated that Boosting produces more diverse ensembles than Bagging and the general trend is that higher diversity usually corresponds to higher accuracy. Diversity measures were used to improve classifier selection by majority voting in [69]. The relationship between diversity measures and ensemble performance for binary classification with simple majority voting using asymmetric misclassification costs was investigated in [52]. It is shown that diversity measure association with ensemble performance is observed using Q-statistics and the coincident failure measure (CFD). Diversity measures were used to guide feature subsets level ensemble in [88, 89], in which the plain disagreement measure (PDM) performed the best. The relationship between different methods of classifier combination and different measures of diversity was studied in [90]. It was found that the Double-Fault measure of diversity and the measure of difficulty both showed reasonable correlation with Majority Vote and Naive Bayes combinations. A survey on various techniques used for creating diverse ensembles and categorizing of them can be found in [91].

However, larger diversity of the ensemble does not always means good performance. Therefore, Kuncheva regarded the diversity as “elusive” [92]. Perhaps diversity is only a necessary condition to the success of the ensemble, there are still some other factors that influence the performance of the ensemble.

3.3 Hybrid Kernel Machine Ensemble

3.3.1 Hybrid Kernel Machine Ensemble

The proposed hybrid kernel machine ensemble (*HKME*) is illustrated in Figure 3.2. A *HKME* consists of two different types of *SVM*s, i.e. a discriminative *BSVC* and a non-discriminative recognition-based ν *SVC* (or *SVDD*). Hence the proposed *HKME* is expected to benefit from the strength of both *BSVC* and ν *SVC*.

HKME is designed for problems where *BSVC* does not perform well or costly

Figure 3.2: The flowchart of hybrid kernel machine ensemble (*HKME*).

to construct while ν SVC shows good performance. For example, there is a kind of imbalanced data problem in which the majority class is compactly clustered and the minority class is scattered in the input space. One example is patient monitoring. The signal morphologies from normal activities (normal class) are similar and the data from this class can be easily collected (majority class), while those from abnormal activities (abnormal class) may exhibit various morphologies and are more difficult to collect (minority class). A discriminative model such as a discriminative *BSVC* can be trained by manually balancing the data or compensating the imbalance using different costs to the two classes. Thus the discriminative model uses the information from both majority class and minority class. However, its performance can still be poor due to the poorly represented minority class. A recognition-based approach such as a one-class *SVM* may do better than the discriminative approach in this situation by modelling the well-represented majority class only. Since the majority class satisfies the assumption of one-class classification where the majority class is well-represented and compactly clustered, it avoids the problem faced by the binary *SVM* due to the inadequate representation of the minority class. However, as a descriptive model, such a recognition-based model is not highly discriminative because the information from the minority class is left totally unused. Hence there is a need to incorporate

the information from the minority class to the recognition-based model or exploit the well-represented majority class further in the discriminative model. Exploiting the complementary nature of these two different types of models, a combination of them is expected to perform better than that of using either of them separately for the classification of this kind of imbalanced data set. Hence constructing a *HKME* by integrating these two hybrid kernel machines in an ensemble is proposed to address this kind of imbalanced data problem. This is the novelty of this proposal.

In this framework, a ν *SVC* can be trained using only the data of majority class, so it can avoid the problem of poor representation of the minority data. On the other hand, a *BSVC* can be trained using balanced data set using oversampling or undersampling, so it benefits from the information from both classes. The outputs of the two *SVMs* can be integrated using some fusion rules. Since the ν *SVC* and *BSVC* are trained using different data sets, the training sets of such two kernel machines can be considered diverse. Furthermore, the different nature of the two *SVMs* can further help to increase the diversity. Therefore, the ensemble of such two kernel machines is expected to improve the classification compared to either of the two types of *SVMs*.

3.3.2 Binary SVM Training

Performance of the classifiers is closely related to the parameters used by the classifiers. There are two hyper-parameters to be tuned in *BSVC* when using Gaussian RBF kernel, the width parameter σ of the RBF kernel and the regularization parameter C which is used to control the tradeoff of errors. The hyper-parameters of *BSVC* can be optimized using cross validation on the training set. The use of cross validation is able to avoid over-fitting [23]. The values of the hyper-parameters are chosen so that the errors of both classes on the validation set are minimized.

Another problem in *BSVC* is the training using imbalanced data sets. It has been shown that balanced data set generally leads to results which are no worse than or superior to those of using natural class distribution, although it does not always produce the optimal results [36]. Since *BSVC* suffers from the imbalanced data problem, the original data set can be balanced first using oversampling, undersampling or *SMOTE* algorithms as introduced in Section 3.2.2. The trained *BSVC* using

the balanced data set can then be integrated with the one-class *SVM* to form the *HKME*.

3.3.3 One-class *SVM* Training

Figure 3.3: (A) Original target data set and (B) generated artificial outliers around the target class in a toy problem in 2-D space.

The hyper-parameters of ν -*SVC* or *SVDD* using Gaussian RBF kernel are the same as those of the BSVCs, i.e. the width parameter of the RBF kernel σ and the regularization parameter ν used to control the tradeoff of errors. The parameters of two-class classifiers can be optimized using cross validation on the training set. However, the information about the outlier class is assumed to be unavailable for one-class classifiers, hence the hyper-parameters can only be estimated using the data from target class or be chosen heuristically.

This problem can be solved by generating artificial outliers [93]. Given a set of target samples, some outlier samples are generated randomly with the assumption that the outliers are uniformly distributed around the target class. The union of targets and generated outliers is used as a validation set to optimize the hyper-parameters of one-class *SVM*. A toy data set and generated artificial outliers are illustrated in Figure 3.3.

As for the imbalanced data problem in question, there are still some outlier samples, i.e., data from the minority class. The hyper-parameters may be tuned to minimize the training error on the whole training set which consists of both major-

ity and minority classes. But this might be dangerous if the minority class is not well represented by the sample data. This problem will be discussed further in the experimental section.

3.3.4 Fusion Rules for Integration of Hybrid SVMs

The integration of two SVMs in a hybrid is a decision level fusion problem. It is nontrivial to properly combine the two sources of information from these two types of SVMs.

Many ensemble learning methods have been developed. In this chapter, several ensemble methods are investigated for the imbalanced data problem, including Decision Template (*DET*), Stacking, Average (*AVG*), Maximum (*MAX*), Minimum (*MIN*), Product (*PROD*) [56, 70] etc.

Let $C_i(x) = \{C_{i1}(x), C_{i2}(x), \dots, C_{ik}(x)\}$ be a set of individual classifiers, called an ensemble, each of which gets an input feature vector $\mathbf{x} = [x_1, x_2, \dots, x_d]^T$ and assigns it to a class label y_i from $Y = \{-1, +1\}$, the goal of the ensemble is to find a class label L_{ens} for $\mathbf{x}$ based on the outputs of k classifiers $C_1(x), C_2(x), \dots, C_k(x)$ corresponding to label $L_1(x), L_2(x), \dots, L_k(x)$. $C_i(x)$ is often an estimate of the posterior probability $P(y_i|\mathbf{x})$.

- Decision template: The decision template DET_j for class $y_j \in \{-1, +1\}$ is the average of the outputs of individual classifiers with respect to the training set for class y_j [56]. The ensemble *DET* assigns the input x with the label given by the individual classifier whose Euclidean distance to the decision template DET_j is the smallest.
- Stacking (Stacked generalization): Taking the output of individual classifiers $C_i(x)$ as input to an upper layer classifier and the final decision is determined by the upper layer classifier [75].

$$L_{ens}(x) = F(C_1(x), C_2(x), \dots, C_k(x)) \quad (3.2)$$

The upper layer classifiers used here include linear discriminant classifiers (*LDCs*) and quadratic discriminant classifiers (*QDCs*) assuming normally distributed

classes. Because the covariance matrices for the classes are near singular, *QDCs* may fail when trying to estimate and invert the covariance matrices [56].

- Average:

$$L_{ens}(x) = \arg \max_j \left(\sum_{i=1}^k \frac{C_{ji}(x)}{k} \right) \quad (3.3)$$

where $j \in \{-1, +1\}$. *AVG* rule calculates the average of the outputs of the k individual classifier and assigns the input x to the class with the largest posterior probability.

- Maximum:

$$L_{ens}(x) = \arg \max_j (\max_i C_{ji}(x)) \quad (3.4)$$

where $j \in \{-1, +1\}$. *MAX* rule takes the maximum value of the outputs from the k individual classifier for each class and assigns the input x to the class with the largest posterior probability.

- Minimum:

$$L_{ens}(x) = \arg \max_j (\min_i C_{ji}(x)) \quad (3.5)$$

where $j \in \{-1, +1\}$. *MIN* rule takes the minimum value of the outputs from the k individual classifier for each class and assigns the input x to the class with the largest posterior probability.

- Product:

$$L_{ens}(x) = \arg \max_j \left(\prod_i C_{ji}(x) \right) \quad (3.6)$$

where $j \in \{-1, +1\}$. *PROD* rule calculates the product value of the outputs from the k individual classifier for each class and assigns the input x to the class with the largest posterior probability.

The problem here is to fuse the outputs of two classifiers. Obviously, the generally used majority voting cannot be used here. Furthermore, it can be proved that Maximum, Minimum, Averaging, Product are equivalent to each other when they are used to combine two classifiers with posterior probability output for a two-class classification task. It has been proved that Maximum and Minimum are equivalent

when combining multiple classifiers for two class classification in [90]. The equivalence of Maximum, Minimum, Averaging, Product when combining two classifiers with posterior probability output for two class classification is proved below.

- Proposition 1. Let $C = \{C_1, C_2\}$ be two classifiers in a two class classification problem $\Omega = \{\omega_+, \omega_-\}$. Let p_1, p_2 be the outputs of the classifiers for class ω_+ and $1 - p_1, 1 - p_2$ be the outputs of the classifiers for class ω_- , where $p_i \in [0, 1], i = 1, 2$. Then the class label assigned to any input x by the Maximum, Minimum, Averaging and Product fusion rules will be the same.
- Proof. It can be assumed that $p_1 > p_2$ without loss of generality. Let $P(\omega_+|x)$ and $P(\omega_-|x)$ denote the posterior probability to the two classes by the ensemble respectively. Consider the two possible relationships between p_1 and $1 - p_2$:

1. $p_1 \geq 1 - p_2 \implies p_1 + p_2 \geq 1$:

(a) Maximum rule: $P(\omega_+|x) = \max \{p_1, p_2\} = p_1$,

$$P(\omega_-|x) = \max \{1 - p_1, 1 - p_2\} = 1 - p_2.$$

So $P(\omega_+|x) \geq P(\omega_-|x)$, Maximum rule will choose ω_+ .

(b) Minimum rule: $P(\omega_+|x) = \min \{p_1, p_2\} = p_2$,

$$P(\omega_-|x) = \min \{1 - p_1, 1 - p_2\} = 1 - p_1.$$

So $P(\omega_+|x) \geq P(\omega_-|x)$, Minimum rule will choose ω_+ .

(c) Average rule: $P(\omega_+|x) = \frac{p_1 + p_2}{2}$,

$$P(\omega_-|x) = \frac{(1 - p_1) + (1 - p_2)}{2} = 1 - \frac{p_1 + p_2}{2}.$$

$$\text{Thus } P(\omega_+|x) - P(\omega_-|x) = \frac{p_1 + p_2}{2} - (1 - \frac{p_1 + p_2}{2}) = p_1 + p_2 - 1 \geq 0.$$

$P(\omega_+|x) \geq P(\omega_-|x)$ holds. So Average rule will choose ω_+ .

(d) Product rule: $P(\omega_+|x) = p_1 p_2$,

$$P(\omega_-|x) = (1 - p_1)(1 - p_2) = 1 - (p_1 + p_2) + p_1 p_2.$$

$$\text{Thus } P(\omega_+|x) - P(\omega_-|x) = p_1 p_2 - (1 - (p_1 + p_2) + p_1 p_2) = p_1 + p_2 - 1 \geq 0.$$

$P(\omega_+|x) \geq P(\omega_-|x)$ holds. So Product rule will choose ω_+ .

Therefore, all the four fusion rules will choose ω_+ when $p_1 \geq 1 - p_2$.

2. $p_1 < 1 - p_2 \implies p_1 + p_2 < 1$:

(a) Maximum rule: $P(\omega_+|x) = \max \{p_1, p_2\} = p_1$,

$$P(\omega_-|x) = \max \{1 - p_1, 1 - p_2\} = 1 - p_2.$$

So $P(\omega_+|x) < P(\omega_-|x)$, Maximum rule will choose ω_- .

$$(b) \text{ Minimum rule: } P(\omega_+|x) = \min \{p_1, p_2\} = p_2,$$

$$P(\omega_-|x) = \min \{1 - p_1, 1 - p_2\} = 1 - p_1.$$

So $P(\omega_+|x) < P(\omega_-|x)$, Minimum rule will choose ω_- .

$$(c) \text{ Average rule: } P(\omega_+|x) = \frac{p_1+p_2}{2},$$

$$P(\omega_-|x) = \frac{(1-p_1)+(1-p_2)}{2} = 1 - \frac{p_1+p_2}{2}.$$

$$\text{Thus } P(\omega_+|x) - P(\omega_-|x) = \frac{p_1+p_2}{2} - (1 - \frac{p_1+p_2}{2}) = p_1 + p_2 - 1 < 0.$$

$P(\omega_+|x) < P(\omega_-|x)$ holds. So Average rule will choose ω_- .

$$(d) \text{ Product rule: } P(\omega_+|x) = p_1p_2,$$

$$P(\omega_-|x) = (1 - p_1)(1 - p_2) = 1 - (p_1 + p_2) + p_1p_2.$$

$$\text{Thus } P(\omega_+|x) - P(\omega_-|x) = p_1p_2 - (1 - (p_1 + p_2) + p_1p_2) = p_1 + p_2 - 1 < 0.$$

$P(\omega_+|x) < P(\omega_-|x)$ holds. So Product rule will choose ω_- .

Therefore, all the four fusion rules will choose ω_- when $p_1 < 1 - p_2$.

Hence all the four fusion rules choose the same label in all the conditions and thus they are equivalent to each other for the two-class problem using two classifiers with posterior probability as outputs.

Due to the equivalence of *MAX*, *MIN*, *AVG* and *PROD* for the two-class problem using two classifiers with posterior probability as outputs, only *AVG* is evaluated in this chapter.

3.3.5 Estimating the Posterior Probability for Outputs of SVMs

The outputs of *SVMs* in equation (2.6) and (2.48) are not posterior probabilities and are in different ranges, and hence are not comparable directly. Thus, their outputs have to be normalized for integration. It is observed that the outputs of *SVMs* show similar forms. One can estimate the posterior probabilities $P_i(y_j|\mathbf{x})$ of the i -th *SVM* using a sigmoid function by minimizing the negative log likelihood of the training data [94]

$$P_i(y_j|\mathbf{x}) = \frac{1}{1 + e^{p_i f_i(x) + q_i}} \quad (3.7)$$

where p_i is a coefficient to control the shape of sigmoid function and q_i is a coefficient to control the shift along the horizontal axis ($f_i(x)$). Thus the ensembles can be constructed using these estimated posterior probabilities.

When estimating the posterior probability of *BSVC*, the training set of the *BSVC* has to be balanced. Otherwise the fitting a sigmoid to the training set of non-linear *SVMs* may lead to biased fits [22, 94]. The balancing of the training set can be implemented using oversampling such as *SMOTE* [33].

To my best knowledge, this is the first attempt to estimating the posterior probability of ν *SVC* or *SVDD*. Since there is only the target data for training a one-class *SVM*, a set of artificial data can be generated whose sample size is the same as that of targets [93]. The union of target data and artificial data can be used to estimate the posterior probability of output from one-class *SVC*.

The posterior probability generated here is only an estimation of the true posterior probability. Bias is unavoidable. This may create some problems to the fusion rules such as *MAX* or *MIN*. However, it may be useful for fusion rules such as stacking or *DET* in which estimating the posterior probability of *SVM* output is used for normalization only.

3.4 Experimental Results and Discussions

The performance of the proposed *HKME* is evaluated on some artificial data sets and real data sets. The evaluation measure is the same as the previous section.

3.4.1 Artificial Data

In order to show the properties of the proposed hybrid kernel machine ensembles, experiments were conducted using a checkerboard data set as described in section 2.5. The negative samples (majority class) lie in half of the checkerboard and the positive samples (minority) lie in a 2×2 square around the negative samples (shown in Figure 2.6). The data are uniformly distributed. The distribution of the checkerboard data is in agreement to the assumption that the negative class is compactly clustered and the minority class is scattered in the input space.

Figure 3.4: The comparison of different schemes using 2×2 checkerboard data set with different imbalance ratio.

The proposed *HKME* is compared to the other generally used method to deal with class imbalance in the 2×2 data set, including oversampling, down-sampling, *SMOTE* and *BSVC* using different costs to the two classes. The number of negative data was fixed as 256, the number of positive data were decreased so that the imbalance ratio (negative to positive) is increased from 1 : 1 to 32 : 1. When the imbalance ratio increases to 32 : 1, the number of positive samples is only 8. The positive data has been too sparse to represent their true distribution. It is thus meaningless to decrease the number of positive data further. The number of test data consists of 1000 positive data and 1000 negative data. The experiment was repeated 10 times and the average value of the *BCRs* achieved by different schemes are reported in Figure 3.4.

The other approaches for comparison include:

- Oversampling: The positive class was randomly oversampled (duplication) so that the training set is balanced.
- Undersampling: The negative class was randomly under-sampled so that the training set is balanced.
- SMOTE: A balanced data set was made by adding some artificially generated

data in between the 3 nearest neighbors of each data in the original data set.

- Using different costs to two classes: The *BSVC* was trained using different costs to the two classes. The primal problem in (2.11) becomes

$$\min_{w,b,\theta} \left\{ \frac{1}{2} \|w\|^2 + C_+ \sum_{i=1}^{N_+} \theta_{i+} + C_- \sum_{i=1}^{N_-} \theta_{i-} \right\} \quad (3.8)$$

where N_+ and N_- are the numbers of positive and negative samples respectively ($N_+ < N_-$) and θ_{i+} and θ_{i-} are the errors of positive and negative samples respectively. The regularization parameter is changes to:

$$C_+ = \frac{C}{2N_+}, \quad y_i = +1 \quad (3.9)$$

and

$$C_- = \frac{C}{2N_-}, \quad y_i = -1 \quad (3.10)$$

Hence the error of minority negative class is penalized more than for the majority positive class in order to compensate the class imbalance.

The parameters of all the *BSVC*s are optimized using 3-fold cross validation. The parameters of the ν *SVC* are optimized using artificially generated outlier data aforementioned. The *BCR* achieved by *HKME* using *AVG*, *DET*, *LDC* and *QDC* fusion rules are reported in Figure 3.5.

Figure 3.4 and 3.5 illustrate the performances of different schemes to address the imbalanced data problem on the checkerboard data set when different imbalance ratios were used. They show the performance variation of these schemes on this kind of imbalance data set with respect to the representativeness of the minority data to their true distribution.

It is observed from Figure 3.4 that discriminative *BSVC* (trained using original data set) perform well when the imbalance ratio is not very high, but its performance deteriorates with the increase of imbalance ratio. *HKME* using *AVG* rule performs the best among all the approaches. The *BSVC* using different costs to two classes perform quite well compared original *BSVC*. Undersampling performs better than original *BSVC*, but is outperformed by using different costs. SMOTE performs

Figure 3.5: The comparison of different fusion rules for $HKME$ using 2×2 checkerboard data set with different imbalance ratio.

reasonably well. It is better than both original $BSVC$ and νSVC . Oversampling performs the worst among all the approaches.

Figure 3.6: The decision boundaries of νSVC , $BSVC$, and $HKME$ using 2×2 checkerboard data set (256 negative samples and 16 positive samples).

The good performance of *HKME* may come from that fact that it benefits from the strength of both of its individual classifiers in the ensemble, the discriminative *BSVC* and recognition-based ν *SVC*. This can be explained using their decision boundaries as illustrated in Figure 3.6. ν *SVC* performs well due to its suitability to model compactly clustered target class. But it has to reject some target samples to form a tighter boundary as mentioned in section 2.4, so it tends to push the decision boundary towards the majority negative class. However, discriminative *BSVC* tends to push the decision boundary toward the minority positive class. The ensemble of these two *SVM* tends to compensate these two different trends and strike a compromise. As shown in the figure, the decision boundary of *HKME* is located in between two classifiers, which is closer to the ideal decision boundary (two squares in the checkerboard).

The *HKME* using four different fusion rules are compared in Figure 3.5. *AVG*, *DET* and *LDC* performs well. But *QDC* does not perform well in some cases. This is because the covariance matrices for the classes are nearly singular in these cases, *QDC*s failed when trying to estimate and inverse the covariance matrices [56]. However, it still performs quite well when it is properly trained as shown in Figure 3.5.

The performance of other approaches in Figure 3.4 may also be explained using their decision boundaries on a checkerboard data set with 256 negative samples and 16 positive samples, as shown in Figure 3.7. The *BSVC* tends to push the decision boundary toward the minority class as aforementioned.

Using different costs to the two classes, the decision boundary tends to be closer to the majority negative class as shown in Figure 3.7 (A). So this approach performs better than *BSVC* trained using original data set.

The artificially generated positive data using *SMOTE* seems quite closer to its original distribution in Figure 3.7 (B), which makes the *BSVC* trained using *SMOTE* performs much better than the other approaches. But it subjects to the assumption that the samples between the nearest neighbors of a sample are from the same class.

Due to the duplication of the minority samples in oversampling, the *BSVC* overfits the minority positive class, which is clearly illustrated in Figure 3.7 (C). This leads to poor performance of oversampling shown in Figure 3.4, especially when the imbalance

Figure 3.7: The decision boundary of *BSVC* training using original data set and those of *BSVC*s trained using (A) different costs to two classes, (B) SMOTE, (C) over-sampled data set and (D) under-sampled data set.

ratio is high. Therefore, random oversampling the minority data is not suitable in *BSVC* training for imbalanced data sets.

Undersampling the majority class seems to produce better decision boundary than that using oversampling. But the shape of the decision boundary is quite different from the ideal one as shown in Figure 3.7 (D). This may be because that some useful information is lost when some samples from the majority class are removed from the training set. This detrimental effect is especially obvious when the size of the minority class is very small.

To sum up, *HKME* performs well in the checkerboard data set. SMOTE and using different costs to the two classes seem quite efficient for this data set. Random undersampling is better than the *BSVC* trained using original imbalanced data set. Random oversampling is not suitable for *BSVC* when the imbalance ratio is high.

Table 3.1: Original, training and test data set used in this study.

Domain	Original			Train		Test	
	# of attributes	Majority	Minority	Majority	Minority	Majority	Minority
Breast	10	444	239	344	139	100	100
Pima	8	500	268	400	168	100	100
Glass	9	185	29	175	19	10	10
Blood	4	127	67	117	57	10	10

3.4.2 Real Data Sets

In order to show the performance of the proposed *HKME* on the real data, the following experiments were conducted using 4 real data sets. Three of them are from UCI database [95], including the Wisconsin Breast Cancer (Breast), the Pima Indian Diabetes (Pima) and the Glass data set (class 7 is taken as minority class). The other is Blood Disorder data set (Blood) from Biomed dataset in the Statlib data archive [96]. These data sets were splitted into training and test data sets randomly. The majority classes (negative) were used to train νSVC . The number of target data was fixed and the number of minority class (positive) was reduced to change the imbalance ratio. The detail of the training and test data were illustrated in Table 3.1. The experiments were repeated 10 times, the average value and standard deviation are reported. The best *BCR* in each case is emphasized using bold fonts.

3.4.2.1 Breast Cancer Data Set

The imbalance ratio was increased from 1 : 10 to 1 : 50 using Breast Cancer data set. The results using different approaches are reported in Table 3.2 and Figure 3.8.

It is observed that *HKME* with *QDC* rule performs the best in all the cases. It seems that this ensemble benefits from the good performance of νSVC whose *BCR* is around 94%. Fusion rules *DET* and *LDC* performs well, but they are outperformed by the original νSVC . *AVG* rule does not perform well. The training of the ensemble is very important. *DET* and *LDC* can produce linear decision boundary on the decision space of νSVC and *BSVC* while *QDC* can produce non-linear decision boundary. It seems that the decision space may not be linearly separable, so *QDC* performs well while the linear fusion rules like *DET* and *LDC* do not perform well as *QDC*. *AVG* rule only calculate the average of the decisions of νSVC and *BSVC*,

Figure 3.8: The performance of different schemes on Breast Cancer data sets.

Table 3.2: BCR (average $\pm$ standard deviation in percentage) of each scheme using Breast Cancer data set.

Imbalance Ratio	1 : 10	1 : 30	1 : 50
νSVC	94.3 ± 1.8	94.3 ± 1.8	94.3 ± 1.8
$BSVC$	93.1 ± 2.5	85.1 ± 3.2	85.1 ± 3.2
Different Costs	95.6 ± 1.0	92.2 ± 4.5	92.2 ± 4.5
Oversampling	50.2 ± 0.2	50.1 ± 0.2	50.1 ± 0.2
Undersampling	95.3 ± 1.5	92.2 ± 2.4	92.2 ± 2.4
SMOTE	88.0 ± 3.3	77.6 ± 6.1	77.6 ± 6.1
$HKME$ (AVG)	92.8 ± 1.0	90.8 ± 2.9	90.8 ± 2.9
$HKME$ (DET)	94.0 ± 1.5	93.6 ± 1.3	93.6 ± 1.3
$HKME$ (LDC)	93.2 ± 1.4	93.2 ± 1.5	93.2 ± 1.5
$HKME$ (QDC)	95.1 ± 1.2	95.0 ± 1.2	95.0 ± 1.2

so its performance is in between νSVC and $BSVC$.

Furthermore, the performance of ensemble rule QDC , DET , LDC and AVG is quite steady with respect to the increase of the imbalance ratio. This shows their ability to deal with this type of imbalanced data problem.

Among the other four approaches to deal with class imbalance, undersampling and using different costs to the two classes perform well. Similar to checkerboard data set, Oversampling almost completely fail due to its overfitting on the minority class. SMOTE does not perform well as that of in checkerboard data set. This may be resulted by the inadequate representation of the minority class. When the number

Figure 3.9: The performance of different schemes on Pima data sets.

Table 3.3: BCR (average $\pm$ standard deviation in percentage) of each scheme using Pima data set.

Imbalance Ratio	1 : 10	1 : 30	1 : 50
νSVC	55.5 $\pm$ 1.5	56.3 $\pm$ 1.6	56.3 $\pm$ 1.6
$BSVC$	57.5 $\pm$ 1.5	53.3 $\pm$ 2.2	53.3 $\pm$ 2.2
Different Costs	73.2 $\pm$ 3.2	71.6 $\pm$ 2.7	71.6 $\pm$ 2.7
Oversampling	50.0 $\pm$ 0.2	50.0 $\pm$ 0	50.0 $\pm$ 0
Undersampling	69.9 $\pm$ 6.5	69.6 $\pm$ 3.8	69.6 $\pm$ 3.8
SMOTE	53.2 $\pm$ 2.4	52.2 $\pm$ 3.0	52.2 $\pm$ 3.0
$HKME$ (AVG)	52.4 $\pm$ 1.4	51.3 $\pm$ 1.0	51.3 $\pm$ 1.0
$HKME$ (DET)	54.6 $\pm$ 2.0	54.5 $\pm$ 1.8	54.5 $\pm$ 1.8
$HKME$ (LDC)	64.9 $\pm$ 8.2	61.7 $\pm$ 7.5	61.7 $\pm$ 7.5
$HKME$ (QDC)	63.1 $\pm$ 6.6	60.8 $\pm$ 8.3	60.8 $\pm$ 8.3

of minority data is very small, the minority data cannot represent its true distribution at all. Even the artificially generated data in between the nearest neighbors cannot add sufficient information about the true distribution of this class. This may be the reason of poor performance of this approach.

3.4.2.2 Pima Data Set

The imbalance ratio was increased from 1 : 10 to 1 : 50 using Pima data set. The results using different approaches are reported in Table 3.3 and Figure 3.9.

It is observed that both νSVC and $BSVC$ perform poorly in this data set.

Figure 3.10: The performance of different schemes on Glass data sets.

HKME with *AVG* and *DET* also perform poorly. However, *HKME* with *LDC* and *QDC* perform well whose *BCRs* are improved around 7% compared to ν *SVC* and *BSVC*. It seems that properly training of the ensemble is very important.

The best approach in this data set is using different costs to the two classes. Undersampling still performs well. SMOTE also performs poorly. Oversampling fail completely.

3.4.2.3 Glass Data Set

The imbalance ratio was increased from 1 : 10 to 1 : 30 using Glass data set. The results using different approaches are reported in Table 3.4 and Figure 3.10.

An interesting observation here is that ν *SVC* was outperformed by *BSVC* in this data set though the imbalanced rate is high. Even so, *HKME* with *LDC* performs well and it has achieved about 3% performance improvement in terms of *BCR* over both ν *SVC* and *BSVC*. But the other *HKMEs* did not perform well. This may be resulted by the fact that the assumption of *HKME* may not be satisfied in this data set.

Similar as in the Pima data set, undersampling and using different costs to the two classes perform well, while oversampling and SMOTE perform poorly.

Table 3.4: BCR (average $\pm$ standard deviation in percentage) of each scheme using Glass data set.

Imbalance Ratio	1 : 10	1 : 20	1 : 30
νSVC	86.0 ± 6.1	86.0 ± 7.7	86.0 ± 7.7
$BSVC$	91.5 ± 3.4	88.0 ± 6.3	88.0 ± 6.3
Different Costs	92.0 ± 3.5	90.5 ± 6.4	90.5 ± 6.4
Oversampling	59.5 ± 5.5	52.5 ± 3.5	52.5 ± 3.5
Undersampling	90.0 ± 7.1	89.0 ± 5.7	89.0 ± 5.7
SMOTE	88.0 ± 4.8	84.0 ± 8.4	84.0 ± 8.4
$HKME$ (AVG)	87.5 ± 3.5	83.5 ± 11.3	83.5 ± 11.3
$HKME$ (DET)	66.0 ± 7.4	68.0 ± 8.6	68.0 ± 8.6
$HKME$ (LDC)	95.0 ± 2.4	91.0 ± 8.1	91.0 ± 8.1
$HKME$ (QDC)	86.0 ± 16.5	72.0 ± 18.3	72.0 ± 18.3

Figure 3.11: The performance of different schemes on Blood data sets.

3.4.2.4 Blood Data Set

The imbalance ratio was increased from 1 : 5 to 1 : 20 using Blood data set. The results using different approaches are reported in Table 3.5 and Figure 3.11.

It is observed that all the $HKME$ s performs well in this data set and show performance improvement over both νSVC and $BSVC$ in all the cases, among which LDC fusion rule perform the best. The reason may be that the distribution of the data in this data set is closely in agreement to the assumption in $HKME$. In this data set, the majority class is the observations made on normal healthy patients while the minority class is those that exhibiting abnormalities due to a rare genetic

Table 3.5: BCR (average $\pm$ standard deviation in percentage) of each scheme using Blood data set.

Imbalance Ratio	1 : 5	1 : 10	1 : 20
νSVC	82.0 ± 9.8	77.5 ± 6.8	77.5 ± 6.8
$BSVC$	77.0 ± 10.6	71.5 ± 8.2	71.5 ± 8.2
Different Costs	86.0 ± 9.7	80.0 ± 12.2	80.0 ± 12.2
Oversampling	59.5 ± 12.3	52.0 ± 6.3	52.0 ± 6.3
Undersampling	82.0 ± 12.3	84.5 ± 9.6	84.5 ± 9.6
SMOTE	75.5 ± 10.4	72.0 ± 14.6	72.0 ± 14.6
$HKME$ (AVG)	85.5 ± 12.3	82.0 ± 10.1	82.0 ± 10.1
$HKME$ (DET)	85.5 ± 8.6	82.5 ± 8.2	82.5 ± 8.2
$HKME$ (LDC)	84.5 ± 8.3	84.0 ± 8.4	84.0 ± 8.4
$HKME$ (QDC)	83.5 ± 11.0	82.0 ± 7.9	82.0 ± 7.9

disease [96].

Using different costs to the two classes and undersampling perform well in this data set. Oversampling and SMOTE perform poorly here.

3.4.3 Comparison of $HKME$ and $1.5SVM$

It has been mentioned at the beginning of this chapter that $HKME$ is a learning machine that is in between one-class classifier and two-class classifier. So $HKME$ is an one-and-half classifier. In fact, $SVDD$ or νSVC trained with the data from two classes can also be called one-and-half classifiers [6, 54]. To some extent, $BSVC$ using different costs to two classes may also be called a type of one-and-half classifier. These approaches were tested using a 2×2 checkerboard data set. The number of majority negative data was fixed as 128 and the number of the positive data was decreased. The number of test data consists of 1000 positive data and 1000 negative data. $BSVC$ s with RBF kernel were trained using these data. The hyper-parameters of the $BSVC$ s were optimized using 3-fold cross validation on the training set. The experiment was repeated 10 times and the average value and standard deviation of the $BCRs$ achieved by $BSVC$ s are reported in Figure 3.12 and Table 3.6.

It is observed in Figure 3.12 and Table 3.6 that the performance of $SVDD$ trained with outliers are quite similar to that of $HKME$. The reason might be that $SVDD$ trained with the data are based on one-class classifiers and it tries to incorporate the information from the other class to the classification. $HKME$ tries to incorporate

Figure 3.12: The performance of different schemes on 2×2 checkerboard data sets.Table 3.6: BCR (average $\pm$ standard deviation in percentage) of each approach using 2×2 checkerboard data set.

Imbalance Ratio	1 : 16	1 : 4	1 : 1
ν SVC	88.5 ± 3.4	92.4 ± 2.2	92.4 ± 2.2
BSVC	81.6 ± 7.1	93.3 ± 0.7	93.3 ± 0.7
BSVC using different costs	87.9 ± 8.5	93.8 ± 2.6	93.8 ± 2.6
SVDD using outliers	95.6 ± 0.8	94.7 ± 1.1	94.7 ± 1.1
HKME (AVG)	96.0 ± 1.1	95.8 ± 0.7	95.8 ± 0.6

the information from the other class by an ensemble of a *BSVC* with a ν *SVC*. So their performances are quite similar to each other.

Furthermore, it is observed that *BSVC* using different costs performs similar to the original *BSVC*. This may be because it is still a two-class classifier though it tends to push the decision boundary to the majority class.

3.4.4 Summary

The proposed *HKME* is evaluated with an artificial checkerboard data set and 4 real data sets. *HKME* performs reasonably well in the artificial data set and most of the real data sets. *HKME* using all 4 fusion rules perform well in the checkerboard data set. But *HKME* using *AVG* and *DET* fusion rules do not perform well in some real data sets. It seems that the optimal decision in the decision space of ν *SVC* and *BSVC* may not be at the mid-point between them. A better way may be to learn

it using stacking such as *LDC* or *QDC*. However, the performance of the stacking depends on its training set and *LDC* or *QDC* may not be the best choice for stacking.

Furthermore, as an one-and-half classifier, *HKME* performs similarly as *1.5SVM*. Its good performance in this kind of imbalanced data sets is based on the ensemble of the two different types of machines.

Among four generally used approaches for class imbalance using *BSVC* as classifier, using different costs to the two classes usually shows good performance. Undersampling perform reasonably well in most cases, but it may be in danger when the number of minority samples is very small. SMOTE performs well only when the number of minority class is not very small and its assumption to generate artificial data is fulfilled. Oversampling by randomly duplicating the minority data tends to overfit the minority class, it is not suitable for *BSVC* in imbalanced data set.

3.5 Concluding Remarks

In this chapter, a new learning algorithms *HKME* is proposed and applied to a kind of imbalanced data problem. It is assumed that the majority class is compactly clustered while the minority class scattered in the input space. Benefited from its two components, discriminative *BSVC* and recognition-based ν *SVC*, *HKME* is shown to perform better than each of them on this kind of imbalanced data sets.

The *HKME* is evaluated using an artificially generated checkerboard data set and 4 real data sets and it shows better or comparable performance compared to the other approaches that are designed to deal with imbalanced data problem. The decision boundary of the *HKME* is usually in between the *BSVC* and ν *SVC* and its property is in between two-class classifiers and one-class classifiers. So it is an one-and-half classifier whose performance is similar to the *1.5SVM*. The *HKME* with stacking rule shows better performance among all the fusion rules in these imbalanced data sets. It seems that *HKME* performs well on this type of imbalanced data sets when its assumption is satisfied and the ensemble is properly constructed. Therefore, the *HKME* is also evaluated in two medical applications in the next two chapters.

Chapter 4

Abnormal ECG Beat Annotation for Long-term Monitoring of Heart Patients

4.1 Introduction

In this chapter, application of one-class classifier and HKME are reported for abnormal Electrocardiogram (ECG) beat annotation for long-term monitoring of heart patients. ECG is a recording of the heart's electrical activity obtained from electrodes attached on the body surface of a patient [97]. Different segments of the ECG signal characterize different cardiac activities. For example, QRS complex reveals the left and right ventricular depolarization, ST-T wave signifies ventricular repolarization, P wave indicates the sequential activation (depolarization) of the left and right atria, etc. A typical normal ECG beat is illustrated in Figure 4.1.

The analysis of heart beat cycles in ECG signal is very important for long-term monitoring and diagnosis of the patients' heart conditions in an intensive care unit or at patients' homes through a telemedicine network. The main purpose of long-term ECG monitoring for the patients suffering from or suspected to suffer from cardiovascular diseases includes:

- Monitoring the effectiveness of treatment for irregular heart rhythms (such as using medication or a pacemaker or automatic defibrillator).

- Evaluating symptoms of possible heart disease (such as chest pain, dizziness or fainting).
- Detecting poor blood flow to heart muscle (ischemia) which may indicate coronary artery disease.
- Detecting arrhythmias that occur intermittently or during certain physical activities.

However, it is very costly for the physicians to analyze the ECG recordings beat by beat since the ECG recordings may last for hours. Therefore, it is significant to develop a computer-assisted technique to examine and annotate the ECG recordings automatically, so to facilitate review by medical experts. This computer annotation will assist physicians to select only the informative (abnormal) beats for further analysis.

Figure 4.1: A typical normal ECG beat.

There are many kinds of abnormal ECG beats corresponding to different cardiac diseases, such as atrial premature beats, ventricular escape beats, left bundle branch block beats and supraventricular premature or ectopic beats, etc. Some of the typical abnormal beats and a normal ECG beat are illustrated in Figure 4.2.

Annotation of ECG recordings requires the discrimination of various types of abnormal ECG beats (abnormal class) from the normal ones (normal class). This is

Figure 4.2: Examples of ECG beats using data from MIT/BIH arrhythmia database. A is a normal beat. The others are abnormal.

a pattern recognition task: one tries to infer a classifier from a limited set of training samples (annotated ECG beats by cardiologist experts), with the assumption that the underlying distribution of these training samples is similar to the true distribution of

the two classes. The trained classifier is then used to annotate the ECG recordings from the other patients. The training data set for the classifier, such as a binary Support Vector Machine (*SVM*), is usually from a large database consisting of ECG records from a large pool of patients.

One of the challenges faced by the ECG beat classifiers is the large variation in the morphologies of ECG signals from different patients. The “normal” range is usually different among human-beings which implies that the distribution of the test data is not the same as that of the training data. Therefore a finely-trained classifier using a large database consisting of the ECG beats from a group of patients may perform poorly when used to annotate the ECG beats from a patient outside of the training database. This is known as the problem of poor generalization. Indeed, even physicians may experience difficulty in assessing abnormal ECG beats if only considering the reference value based on the general patient population. Hence it is necessary to incorporate local information of each patient to improve the annotation of abnormal ECG beats and thus help to improve the generalization.

In this chapter, a new concept learning-based patient-adaptable ECG beat annotator (4.4) is proposed to address such a generalization problem. A concept learning model, νSVC [4] (Section 2.4.2) is trained, using only about 5 minutes of normal ECG beats from a patient, to learn the concept “normal”. The learned model is then used to determine whether a new ECG beat belongs to the concept “normal” or not, thus to annotate abnormal beats in the long-term ECG recording of the same patient. The main advantages of the proposed concept learning approach includes:

- Good generalization: Its generalization is better than those of the classifiers trained using a lot of ECG beats from a large group of patients.
- Convenience: The proposed approach can relieve the physicians from annotating the ECG beats for training a patient-adaptable local ECG annotator.
- Less computation complexity: Trained using only about 5 minutes of normal ECG beats from each patient, the learned νSVC model is more compact compared to the binary *SVMs* trained using a lot of ECG beats from a large group of patients. Hence it is faster in annotating the heart beats in the long-term ECG recordings.

The proposed concept learning-based patient adaptable ECG annotator was evaluated using MIT/BIH arrhythmia database [9]. Experimental results show that the proposed concept learning-based patient adaptable ECG annotator constructed using only hundreds of normal ECG beats from a patient outperforms the other classifiers trained using tens of thousands of data from a large group of patients in annotating the abnormal ECG beats of the same patient.

Although the specific local information from each patient is significant in annotating the ECG beats from the same patient, the standard reference value from a large group of patients is also useful in evaluating the degree of “normal” of the ECG beats from each patient in clinical application. Motivated by this, a hybrid kernel machine ensemble is proposed to combine the information from the specific patient and that from a large group of patients. The hybrid kernel machine ensemble consists of two base annotators, a νSVC and a $BSVC$. In the proposed framework, a νSVC is constructed using only 5 minutes of normal ECG beats from each patient to adapt to the patient so that this local νSVC represents the specific reference value of the patient. Furthermore, a binary SVM ($BSVC$) is constructed using many ECG beats from a large group of patients so that this global $BSVC$ represents the standard reference value of this large group of patients. These two types of SVM s perform differently in annotating the ECG beats from the same patient due to the different information they have incorporated. Integration of these two ECG annotators using an ensemble is expected to perform better than using each of them separately. Experimental results using MIT/BIH arrhythmia database show that the proposed patient-adaptable hybrid kernel machine ensemble outperforms both the global $BSVC$ and the local νSVC , thus supporting its feasibility in practical clinical application.

4.2 Generalization and Imbalanced Data Problem in ECG Beat Annotation

4.2.1 Generalization Problem

A fundamental assumption in the field of pattern recognition is that the underlying distribution of the training samples is the same as that of the test samples. However,

such assumption may not hold in practical application. The abnormal ECG beat annotation problem is one of the examples. Figure 4.3 illustrates the distribution of the first two principal components of the original 181-dimensional(D) feature vector of ECG beats obtained by using Karhunen-Loeve transform (PCA) from 8 recordings of MIT/BIH arrhythmia database (refer to Section 4.4.2.1 for detail), preserving 69% of the total variance, where the circles indicate normal ECG beats and the cross signs are abnormal ones. Although some discriminative information may be lost using PCA, it can be observed that the distributions of “normal” ECG beats are different among patients. Figure 4.4 illustrates the distribution of the ECG data from 44 recordings of MIT/BIH arrhythmia database using same PCA projection as in Figure 4.3. The difference between the ECG data from each patient in Figure 4.3 and the ECG data from 44 patients in Figure 4.4 is quite prominent. This is the difference between the population and each patient. Although an ECG detector can be finely trained using the ECG beats from a large database which consists of the ECG beats from different patients, it may perform poorly in annotating the ECG beats of other patients who are not in the database. This is the problem of poor generalization.

The solution to such generalization problem lies in the incorporation of local information of a specific patient to the ECG annotator. Since the distribution of the training samples is not the same as that of the test samples, some information about the true distribution of samples from each patient has to be added to train the ECG annotator properly.

4.2.2 Imbalanced Data Problem

In the scenario of long-term monitoring of some patients suffering from cardiovascular diseases, the normal ECG beats usually dominate the ECG recordings such as in patients suffering from or suspected to suffer from asymptomatic heart failure, congestive heart failure, cardiac dysfunction and cardiac arrhythmias etc [98], i.e. the number of abnormal ECG beats is far less than that of the normal ones. It may take a long time to collect sufficient and balanced normal and abnormal ECG data to construct a good classifier otherwise the classifier may suffer from the imbalanced data problem [21] aforementioned in chapter 2.

Figure 4.3: Scatterplot of ECG data of 8 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection.

4.3 Related Work

The ECG beat annotation problem at hand can be considered as a two class pattern recognition problem. Many algorithms have been applied to ECG beat cycle analysis,

Figure 4.4: Scatterplot of ECG data of all the ECG data from 44 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection.

which can be categorized into two broad classes, i.e. knowledge-based approach and statistical learning-based approach.

The knowledge-based approach tries to mimic the diagnostic process of cardiologists in annotating the ECG beats by modelling the knowledge of the cardiologists using a knowledge-base. The knowledge-base can be set up in consultation with the experts in the field of cardiology [99]. For example, Taddei *et al* developed a knowledge-based system for interpretation of arrhythmias in long-term ECG [100]. A fuzzy pattern matching technique was presented for diagnostic ECG classification in [101]. In order to improve the ECG beat annotation by incorporating the knowledge of cardiologists, fuzzy logic was investigated in [102, 103]. A survey of knowledge-based ECG interpretation can be found in [99]. However, the modelling of the knowledge of the cardiologists is very difficult because the knowledge or expertise is usually fuzzy and hard to quantify which limits its clinical application.

Statistical learning-based approach has been the active area of research compared to the knowledge-based approach for the interpretation of ECG signal. For example, time-domain template matching was used in [104] and [105]. Principal Component Analysis (PCA) of the ECG signal was investigated for cardiac arrhythmia classification in [106]. Wavelet-based Bayesian models were used for ECG beat classification [107]. A linear discriminant classifier was introduced for classification of heartbeats [108].

Among all the techniques used in the classification of ECG beats, neural networks are the most generally used models. Various types of neural networks have been investigated. For example, Kohonen self-organizing maps were investigated in ECG beat recognition in [109–111]. Learning vector quantization was employed in [111–113]. Autoassociators were investigated in [114, 115]. Multi-layer perceptrons were used in [116–119]. A probabilistic neural network array architecture for ECG classification was proposed in [120]. Stamkopoulos *et al* proposed to detect ischemia using non-linear PCA neural networks [115]. A survey of ECG pattern recognition based on nonlinear transformations and neural networks can be found in [121].

There were some attempts to incorporate fuzzy logic into the neural networks for ECG analysis, such as [122–125]. Some attempts were made to construct hybrid neural network based on genetic algorithms [126–128]. Recently, support vector machines, a new method emerged from the neural network research community, were introduced for ECG signal recognition [19, 20, 129]. One classifier may not be good enough to discriminate the ECG beats, hence there were also investigation of ensemble learning for ECG beat classification by combining a set of classifiers such as in [19, 130, 131].

These ECG beat annotation algorithms are usually trained using a large ECG data set from a group of patients. Hence, they suffer from the poor generalization problem mentioned in section 4.2. Some attempts have been made to address such problem. For example, Watrous *et al* proposed a patient-adaptive neural network for heart patient monitoring in which the weights of a multi-layer feed-forward neural network can be changed by a patient model so as to adapt the neural network to each patient [132]. However, it is difficult to set the parameters of the patient model, which limits its utility in real clinical application. Osowski *et al* presented the application of fuzzy self-organizing neural network and higher order statistics for ECG beat classification to address the large variation in the morphologies of ECG waveforms. But it is difficult to choose the optimal number of clusters for the fuzzy clustering algorithm. In their inspiring paper, Hu *et al* proposed a mixture of expert approach to address this problem [112, 133, 134]. Such a mixture of expert structure was formed by combining the knowledge of a global expert trained using ECG data from a large database and a local expert trained using several minutes of ECG signals from a specific patient.

When the mixture of expert system was used to classify the ECG beats from the specific patient, classification performance was improved compared to that based on the global expert only. However, the major drawback of such an approach is that some ECG beats from each patient have to be annotated by a physician in order to train the local expert, even with only several minutes of the patient's ECG recording, though the local expert is helpful to improve the generalization of the ECG annotator. Such annotation process is very costly and discourages the practical application of this approach [108, 112].

A novel concept-based approach (section 4.4) and a hybrid kernel machine ensemble approach (section 4.5) have been proposed to address these problems, which are introduced in the following sections.

4.4 Concept Learning-based Approach

4.4.1 Proposed Methodology

The flowchart of the proposed concept learning-based ECG beat annotator is illustrated in Figure 4.5. A concept learning model, νSVC is trained using only about 5 minutes of normal ECG beats from a patient to adapt to the specific reference value of the patient. The trained model can then be used to determine whether the other ECG beats from the same patient belong to the “normal beats”. Hence the abnormal ECG beats can be annotated automatically for further analysis.

As mentioned in section 4.2.1, there is an innate difference between the normal range of each patient and that of a group of patient. There is a need to incorporate the local information of each patient to improve the generalization of the ECG beat annotator. Figure 4.6 and Figure 4.7 illustrates the ECG data of 8 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection, which is the same projection of data from the same patients as in Figure 4.3. Each recording is from a patient which lasts about 30 minutes. The ECG data from the first 5 minutes of each recording (training set –TN) and the last 15 minutes of each recording (test set –TT) are plotted respectively. Compared to the “normal” range of a group of patients in Figure 4.4, the “normal” range of the ECG data in

Figure 4.5: Flowchart of the proposed framework for abnormal ECG beat annotation.

the first 5 minutes of a patient in Figure 4.6 and Figure 4.7 is more similar to that of the last 15 minutes of the same patient. Since the distribution of training data of the νSVC model is more similar to the ECG beats of the same patient than those of the data from a large group of patients, the νSVC model trained properly using about 5 minutes of the “normal” ECG data from a patient is expected to perform better than the classifiers trained using the data from a large group of patients.

The concept learning model, νSVC has been explained in section 2.4.2. The detail of the proposed concept learning-based patient-adaptable ECG beat annotator as shown in the flowchart of Figure 4.5 is introduced in the following section.

4.4.1.1 Signal Preprocessing

During the acquisition of ECG signal, the ECG signal is usually coupled with noise and baseline shift due to power line interference, respiration and muscle tremors etc,

Figure 4.6: Scatterplot of ECG data of 4 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection. “TN” are the first 5 minutes of each recording while “TT” are the last 15 minutes of each recording).

which have to be removed in favor of further analysis. The ECG signal is firstly processed using two averaging filters to suppress the noise [135]:

Figure 4.7: Scatterplot of ECG data of 4 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection. “TN” are the first 5 minutes of each recording while “TT” are the last 15 minutes of each recording.

- Power line interference suppression: averages samples in a period of the interference frequency of the power line.

- Electromyogram noise suppression: averages samples in an interval of 28 ms.

The baseline of the ECG signal can be obtained by applying two consecutive median filters to the ECG signal after noise suppression whose widths are 200ms and 600ms respectively [108]. The baseline is subtracted from the ECG signal after noise suppression and the resulted signal is then baseline-corrected.

After noise suppression and baseline correction, the R-peak of the ECG signal can be detected using the first derivative of the ECG signal as in [135].

4.4.1.2 Feature Extraction

The performance of the ECG annotator is closely related to the features used. The generally used features for ECG beat classification includes time domain representation [108,112], nonlinear dynamic modelling [136,137], heartbeat interval features [108], Hermite functions [138], autoregressive modelling-based representation [139,140], Fourier transform-based representation [141] and wavelet transform-based representation [124,125,129,141], etc. In this study, the raw amplitude of the time domain ECG signals after noise suppression and baseline shift removal was investigated as feature vectors to represent the ECG beats.

After the R-peak is detected as mentioned in the previous section, the ECG signal in a window of 500 ms is taken as an ECG beat. The lengths of the signal before and after the R-peak in each beat are 167 ms and 333 ms respectively, such that the window covers most of the characterization of the ECG beat (for an ECG signal sampled at $360Hz$, 180 samples around each R-peak are taken in a window, with 59 samples before the R-peak and the other 120 samples behind the R-peak). The amplitude of sampled signal in each window are then taken to form a feature vector of 180-dimensions. It has been shown that R-R interval (the interval between two consecutive R-peaks) is useful in recognition of some abnormal ECG beats [108,112]. Therefore, it is also included in this study by appending it to the 180-dimensional feature vector. The length of the feature vector to represent the ECG beat is then 181.

Note that only R-peak detection is needed in signal preprocessing, no QRS complex detection is needed. The exclusion of the QRS complex detection significantly

relieves the computation complexity of the signal processing. This is one of the advantages of the proposed methodology.

4.4.1.3 Normalization

There are some variation in the amplitude ranges of ECG signals among the human-beings. Hence a normalization procedure to the ECG feature vectors is necessary otherwise the ECG beats may not be comparable. The feature vectors are divided by the mean value of R-peaks in the training data of each patient, such that the maximum amplitude in each ECG beat window is around 1. The normalized ECG feature vectors are then used for the annotation process using the trained concept learning models.

4.4.1.4 ν SVC-based Patient-adaptable ECG Beat Annotation

Using a short period of the normal ECG beats from a patient to be monitored, a ν SVC model can be trained to learn the concept of “normal” beats of the patient. A new ECG beat can be classified to the “normal” class or non-“normal” (abnormal) class by the trained ν SVC model. The abnormal ECG beats are thus annotated automatically for the physicians for the further review.

It can be observed in Figure 4.6, Figure 4.7 and Figure 4.3 that the normal ECG beats are compactly clustered in the input space while the abnormal beats are usually scattered. This type of data distribution is suitable to be modelled by a one-class classifier as mentioned in Chapter 2 and Chapter 3. It is thus reasonable to generate some artificial outliers (abnormal class) by assuming uniform distribution around the compactly clustered normal class to select hyperparameters for the model as explained in section 3.3.3.

4.4.2 Experiments

The following experiments were conducted using the MIT/BIH arrhythmia database to demonstrate the feasibility of the proposed concept learning-based patient-adaptable ECG beat annotator and evaluate its performance.

4.4.2.1 Experimental Setting

The MIT/BIH arrhythmia database is a benchmark database for evaluating the ECG beat annotators. It consists of 48 annotated recordings from 47 patients and each recording is about 30 minutes in length. The labels in the annotation files made by cardiologists are used as the ground truth in training and evaluation of the ECG beat annotators. In this study, the ECG beats labelled as “normal” (NOR) are taken as the target class whose number in the database is about 70000. All of the other beats are regarded as “abnormal” class, including nodal premature beats, atrial premature beats and ventricular escape beats etc. whose number is around 30000.

Four recordings (#102, 104, 107, and 217) including paced beats are excluded from the study in compliance to the standards recommended for reporting performance results of cardiac rhythms by the Association for the Advancement of Medical Instrumentation (AAMI) [142]. The remaining 44 recordings are divided into two categories.

1. 18 recordings are selected as 18 local training sets and test sets, in which the number of abnormal beats is significantly less than that of the normal beats, which is in agreement to the scenario in long-term monitoring of patients suffering from cardiovascular diseases. These recordings include # 100, 105, 106, 119, 200, 203, 205, 209, 210, 213, 215, 219, 221, 223, 228, 230, 233 and 234. Each of the 18 recordings is split into two sets.
 - The normal ECG beats in the first 1/6 of each of the 18 recordings (about 5 minutes or 300 beats) are used as the local training set to construct the νSVC s.
 - The last 15 minutes of each of the 18 recordings (about 1000 beats) is used as test set to evaluate the performance of the ECG annotators.

The detail of the number of normal and abnormal ECG beats in each recording is tabulated in Table 4.1.

2. 10000 ECG beats (with half normal beats and half abnormal beats) from 26 recordings are used as global training set (DB_G) to train some classical binary

classifiers for comparison with the proposed concept learning-based patient-adaptable ECG beat annotator. These recordings are # 101, 103, 108, 109, 111, 112, 113, 114, 115, 116, 117, 118, 121, 122, 123, 124, 201, 202, 207, 208, 212, 214, 220, 222, 231 and 232.

Table 4.1: Description of the ECG training and Test data sets for reference. * Note that only “Normal” ECG beats are used in training the local ν SVCs

Recording	Training Set		Test Set	
#	Normal	Abnormal *	Normal	Abnormal
100	374	4	1113	22
105	425	23	1241	102
106	278	71	670	377
119	246	102	738	307
200	309	156	802	592
203	428	89	1283	268
205	444	1	1267	67
209	492	16	1398	125
210	409	38	1185	155
213	426	122	1361	284
215	524	42	1595	103
219	368	17	983	171
221	328	82	1055	174
223	402	38	946	373
228	267	89	882	186
230	377	33	1160	71
233	363	162	1124	450
234	460	0	1373	7
Total	6920	1085	20176	3834
Average	384	60	1121	213

4.4.2.2 Training Global Classifiers for Comparison

In order to compare with the proposed concept learning-based patient-adaptable ECG beat annotator, some generally used binary classifiers are investigated which are trained using the ECG data from a group of patients, DB_G .

1. Normal densities based linear classifier (LDC): The linear classifier between the two classes by assuming normal densities with equal covariance matrices.
2. Normal densities based quadratic classifier (QDC): The quadratic classifier between the two classes by assuming normal densities.

3. Nearest Mean Linear Classifier (NMC): The nearest mean classifier between the two classes. The test pattern is classified to the class whose mean value is closer to the test pattern in the input space.
4. Feed-forward neural network classifier (NN): A feed-forward neural network classifier trained using back-propagation. The parameters are optimized using 3-fold cross validation on the training set.
5. Linear Binary Support Vector Classifier (LSVC) [11]: A binary *SVM* with linear kernel. The regularization parameter are optimized using 3-fold cross validation on the training set.
6. Binary Support Vector Classifier with Radius Basis Function (RBF) kernel (*RSVC*): A binary *SVM* with RBF kernel. The hyperparameters are optimized using 3-fold cross validation on the training set.

A Matlab toolbox, PRTOOLS [143] is used in the experiments to construct LDC, QDC, NMC and NN. LIBSVM [144] is used to construct ν *SVC*, LSVC and *RSVC*.

4.4.2.3 Training Local One-class Classifiers for Comparison

The ν *SVC*-based patient-adaptable ECG beat annotator are compared to some generally used one-class classifiers which are listed below.

1. Nearest Neighbor Data Description (*NNDD*): It can be derived from a local density estimation by the nearest neighbor classifier [51]. The method only uses distances to the first nearest neighbor.
2. Mixture of Gaussians (*MOG*): This method estimate the density of the target class using a mixture of Gaussians, which is a linear combination of normal distributions [23]:

$$p_{MOG} = \frac{1}{M} \sum_{i=1}^M \lambda_i p_N(x; \mu_i, \sigma_i) \quad (4.1)$$

where λ_i s are the mixing coefficients,. The mean μ_i and covariances σ_i of each Gaussian components can be estimated by expectation minimization [23, 51].

A Matlab toolbox, DD-TOOLS [145] is used in the experiments to construct these one-class classifiers.

Table 4.2: Full classification matrix for calculating the evaluation criteria used in this study.

Ground Truth	Classification Results	
	Abnormal	Normal
Abnormal	True Positive	False Negative
Normal	False Positive	Truth Negative

4.4.2.4 Evaluation Measure

The full classification matrix for calculating the evaluation criteria used in this study is illustrated in Table 4.2. Four cases can be observed in the classification results, i.e. true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Three evaluation measures are employed in evaluating the performance of ECG beat annotation, including sensitivity (SEN), specificity (SPE) and balanced classification rate (BCR).

- *Sensitivity* is the fraction of abnormal ECG beats that are correctly detected among all the abnormal ECG beats.

$$SEN = \frac{TP}{TP + FN} \quad (4.2)$$

- *Specificity* is the fraction of normal ECG beats that are correctly classified among all the normal ECG beats.

$$SPE = \frac{TN}{TN + FP} \quad (4.3)$$

- Balanced Classification Rate is the algebraic mean of SEN and SPE.

$$BCR = \frac{1}{2}(SEN + SPE) \quad (4.4)$$

The generally used Average Classification Rate (ACR) is also calculated in the experiments.

- Average Classification Rate is the fraction of all correctly classified ECG beats

- regardless of normal or abnormal among all the ECG beats.

$$ACR = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.5)$$

As mentioned in the previous section, the “normal” class dominates the test set. Commonly used average classification rate is not valid in evaluating the performance of the classifiers in such an imbalanced data set. For example, if a classifier is trained to classify all the test data as normal beats, it has $SPE = 100\%$ and $SEN = 0\%$, but the ACR is still high due to the number of abnormal beats being too small. However, its BCR is only 50%. In this kind of imbalanced data set, similar to the geometric mean of the SEN and SPE used in [30], BCR is more suitable for evaluating the performance of the classifiers than the generally used AVR. Only when both SEN and SPE have large value can BCR have a large value. Therefore, the use of BCR can have a balanced performance in the evaluation of the classifiers in this kind of imbalanced data sets which favors both lower false positives and false negatives.

4.4.2.5 Results and Discussion

The annotation results of using the binary classifiers trained on the large data set DB_G and the νSVC trained using only 5 minutes of “normal” ECG beats from each patient are illustrated in Table 4.3. The reported results are averaged over 18 test ECG recordings. In accordance with AAMI recommendations to present the results by each patient, the annotation results of νSVC and the best binary classifier – $RSVC$ in each recording are shown in Figure 4.8 and Table 4.4. Test sets #1 – #18 correspond to recording 100, 119, 200, 203, 205, 209, 210, 213, 215, 219, 221, 223, 228, 230, 233, 234, 105 and 106 in MIT/BIH arrhythmia database respectively. The number of Support Vectors and the test time of νSVC and $RSVC$ are reported in Table 4.5.

The size of the local training set of νSVC model is important. The size of the local training set of the νSVC model is varied and the relation between the size of the training set of νSVC model and the performance of the νSVC model is reported in Figure 4.9.

Table 4.3: Results (average $\pm$ standard deviation) of abnormal ECG beat annotation (in percentage).

Classifiers	BCR	SEN	SPE	ACR
RSVC	78.1 ± 19.5	82.3 ± 30.9	74.0 ± 30.6	76.3 ± 27.0
LSVC	74.8 ± 14.6	57.2 ± 29.1	92.4 ± 19.2	87.8 ± 15.9
LDC	73.4 ± 13.9	50.2 ± 28.0	96.7 ± 10.2	90.6 ± 9.9
NMC	75.1 ± 15.8	52.3 ± 32.6	98.0 ± 5.2	91.9 ± 6.9
NN	71.5 ± 17.4	51.1 ± 32.9	91.9 ± 22.4	87.2 ± 18.7
ν SVC	91.3 ± 10.9	89.0 ± 22.4	94.2 ± 7.3	94.1 ± 5.8
NNDD	87.5 ± 13.2	81.7 ± 26.4	93.3 ± 7.8	92.2 ± 6.1
MOG	90.7 ± 13.5	88.1 ± 26.2	93.4 ± 7.9	93.3 ± 7.0

Figure 4.8: Comparison of the annotation results of the ν SVC and RSVC in each recording of the test sets in terms of BCR.

- Generalization

It is observed from Table 4.3 that the proposed patient-adaptable concept-learning model, ν SVC outperforms all the other global classifiers trained using the data from 26 patients. The best global annotator is RSVC whose BCR is 78.1%. The proposed patient-adaptable ν SVC-based annotator achieved 91.6% in terms of BCR, which is 13.5% higher than the best global annotator RSVC and with less variance. This result supports the claim that the patient-adaptable concept learning-based approach is better than the classifiers trained using the

Table 4.4: Annotation results of the ν SVC and RSVC in each recording of the test sets.

Recording	RSVC				ν SVC			
#	TP	FN	FP	TN	TP	FN	FP	TN
100	1	21	2	1111	22	0	239	874
105	102	0	262	979	102	0	268	973
106	374	3	69	601	377	0	54	616
119	305	2	0	738	307	0	170	568
200	570	22	49	753	572	20	23	779
203	263	5	889	394	237	31	55	1228
205	58	9	0	1267	67	0	50	1217
209	77	48	924	474	24	101	3	1395
210	133	22	72	1113	152	3	151	1034
213	256	28	479	882	248	36	180	1181
215	1	102	0	1595	103	0	24	1571
219	162	9	932	51	170	1	272	711
221	174	0	132	923	174	0	87	968
223	255	118	423	523	361	12	176	770
228	183	3	196	686	153	33	0	882
230	70	1	852	308	71	0	91	1069
233	442	8	77	1047	445	5	49	1075
234	7	0	1	1372	7	0	111	1262
Total	3433	401	5359	14817	3592	242	2003	18173

Figure 4.9: The relation between the training set size of ν SVC model and its performance in terms of BCR (Mean $\pm$ Standard Deviation). From 1 to 10, the length of the ECG signal used in the training set decreases from 15 (15/1) minutes to 1.5 (15/10) minutes.

Figure 4.10: Scatterplot of ECG data of 2 test sets in MIT/BIH arrhythmia database showing the first two principal components of PCA projection.

data from a large group of patients (DB_G) in the annotation of the ECG beats from each patient. It indicates that the local information of each patient is very important in the classification of the ECG beats from the same patient. The incorporation of such local information is able to fill in the gap between the distribution of training data set and test data set, thus help to improve the generalization of the ECG beat annotator. Therefore, the proposed patient-adaptable concept learning-based approach is suitable to address the problem at hand.

In Figure 4.8, the performance of νSVC varies among the test sets. The νSVC model performs well in 16 among all the 18 test sets which shows its good generalization. Only in 2 among all 18 recordings it was outperformed by the global $RSVC$. The ECG data in these two recordings in Figure 4.10 are plotted using the same PCA projection as used in Figure 4.3. It is observed that the two classes of data are roughly well separated. In this cases, the binary $RSVC$ is good enough. Because some normal ECG beats need to be excluded from the decision boundary of νSVC to form a tighter boundary as mentioned in section 2.4, this resulted in some error when annotating the normal data (false positives). This may be the reason why $RSVC$ outperforms νSVC in these two recordings. However, it should be noticed that the BCR achieved in these two recording by νSVC is still above 95%.

- Comparison with other one-class classifiers

It can be observed in Table 4.3, one-class classifiers *NNDD* and *MOG* also outperforms all the global classifiers. This support the claim that local information is more helpful in classification the abnormal ECG beats of the patients. However, these two one-class classifiers were outperformed by the ν *SVC* model though the difference between ν *SVC* and these two one-class classifiers are much smaller than that of between ν *SVC* and the global binary classifiers.

- The influence of the size of training set to the performance of ν *SVC*

The influence of the number of training samples for the proposed ν *SVC*-based ECG beat annotator to its performance can be observed in Figure 4.9. Obviously, more training samples imply better annotation performance. But the difference of performance between using normal ECG beats of 5 minutes and those of 15 minutes is not very much. Although more training samples is good for better annotation, it also means longer time to construct the ECG annotator. Five minutes of normal ECG data seems to be a good tradeoff between the performance and the convenience which can be observed from the figure.

- Features

The classification results in Table 4.3 and Figure 4.8 show that the features used are quite efficient in discriminating abnormal ECG beats from those of the normal ones. Compared to other features such as heartbeat interval features [108], autoregressive modelling-based representation [139,140] and wavelet transform-based representation [124,125,129,141], etc., the currently used features are simpler to implement. Because only R-peak of each ECG beat needs to be detected to extract these features and no further effort is needed such as QRS detection or other transforms. Therefore, its computation complexity is far less than those of the other features.

- Computational complexity of the annotation

More support vectors means more computation time in the annotation. The average number of support vectors of the ν *SVCs* is compared to that of the binary *RSVC* in the experiment in Table 4.5. The computation time in the table

Table 4.5: Comparative results of testing time and number of Support Vectors (average $\pm$ standard deviation) of abnormal ECG beat annotation. The test time is the total time of annotation of the 18 recordings (24010 beats).

Classifiers	Test time (in seconds)	Average # of Support Vectors
νSVC	0.163	8.3 ± 2.5
$RSVC$	39.352	1641

was calculated on a server with Intel Xeon CPU, 2.8 GHz frequency and 4 GB memory. The average number of support vectors of νSVC model among the 18 ECG recordings is 8.3 while the number of support vectors of the $RSVC$ trained using DB_G is 1641. The decision function of $RSVC$ (2.6) has the same form as that of νSVC model (equation 2.48). Therefore, the proposed νSVC -based patient-adaptable ECG annotator is much faster than the global $RSVC$ -based ECG annotator, which makes it easier to implement in real application. It can also be observed by the test time of the two kinds of kernel machines in Table 4.5, where the test time of $RSVC$ is about 200 times of that of νSVC .

- Comparison with the other methods:

Hu *et al* [112] concentrated on the classification of normal beats and ventricular ectopic beats using a mixture of two classifiers. The sensitivity and specificity achieved are 82.6% and 97.1%, which means its BCR is about 90%. de Chazal *et al* [108] have claimed that they achieved comparable performance to the method of Hu's [112] using a linear discriminant classifier. Due to the different experimental settings, the results of their approaches and the proposed approach are not directly comparable. The proposed patient-adaptable concept learning-based approach achieved a BCR of 91.3% although only about 5 minutes of "normal" ECG beats from each patient are used to train the local νSVC model in an unsupervised mode. Furthermore, the data recordings including the test data and the training data of νSVC are seriously imbalanced. The classifiers used in [112] and [108] have problems in training a good classifier in such cases. Therefore, the proposed approach shows at least comparable performance compared to their methods.

Another advantage of the proposed approach is that it can relieve the physicians

from annotating the ECG beats one by one as needed in [112] and it is easier to be constructed to adapt to each patient. Hence it can be applied conveniently in real clinical application.

4.5 Hybrid Kernel Machine Ensemble Approach

4.5.1 Proposed Methodology

Although the concept learning-based approach proposed in the previous section performs well in the abnormal ECG beat annotation, it may be possible to be further improved. When a physician examines the ECG recordings of a patient, the physician considers not only the specific reference from each patient to be examined, but also the standard reference from the patient-group due to the innate difference between the normal range of each patient and that of a group of patient aforementioned. That is to say, the diagnosis made by the physician is based on the information from both the patient-group and each specific patient. Motivated by this, a *HKME*-based approach is proposed for the ECG beat annotation problem for long-term monitoring heart patients.

Figure 4.11 illustrates the flowchart of the proposed Hybrid Kernel Machine Ensemble (*HKME*)-based ECG beat annotator. This *HKME* consists of two base classifiers, one is a binary *SVM* trained using the ECG data from a large group of patients, the other is a concept learning model, ν *SVC* trained using only about 5 minutes of normal ECG beats from each patient to be monitored. The final decision is determined by a fusion rule. The recognition-based ν *SVC* has been described in the previous section. It represents the specific reference value of the patient. The discriminative binary *SVM* is incorporated the global information of a large group of people and thus it can be regarded as the reference values based on the general patient population. Due to different information learned by these two *SVMs*, they usually perform differently in classifying the ECG beats in the long-term ECG recording of the patient. Furthermore, ν *SVC* is a non-discriminative recognition-based model and *BSVC* is a discriminative model. Due to the complementary nature of such two types of *SVMs*, integration of the two types of kernel machines using an ensemble is

Figure 4.11: Flowchart of the proposed framework for abnormal ECG beat annotation.

expected to perform better than using either of them separately.

Of course, the difference between the specific reference value of each patient and the standard reference value of a patient-group varies among patients. Therefore, the dependence of the final decision on these two kinds of reference values should be varied too. As for the *HKME*, the ensemble of the two types of *SVMs* should be varied with respect to each patient. So the performance of the *HKME* is expected to vary among patients as well.

The concept learning model, νSVC has been introduced in section 4.4. The detail of the proposed *HKME*-based patient-adaptable ECG beat annotator as shown in the flowchart of Figure 4.11 is introduced in the following section.

4.5.1.1 Signal Preprocessing, Feature Extraction and Normalization

The signal processing, feature extraction and normalization of the ECG signal are the same as the concept learning-based approach in section 4.4.1.

4.5.1.2 HKME-based Patient-adaptable ECG Beat Annotation

In the training stage in Figure 4.11, the two *SVMs* are trained separately. The global binary *SVM* can be properly trained using the ECG data from a large patient-group. For each patient to be monitored, a ν *SVC* model can be trained using a short period of the normal ECG beats from the patient. The fusion rule can be adapted using the normal ECG beats from the patient and some artificial data generated as reported in section 4.4.1.4.

In the monitoring stage, the ECG recording of each patient can be processed firstly following the procedure as described in the previous section to obtain the feature vectors representing the ECG beats. The feature vectors are then classified as normal or abnormal by the two *SVMs* firstly. The final annotation decision is made by the fusion rule. The annotated abnormal ECG beats are then selected automatically to the physicians for further review.

The problem at hand can be considered as a data fusion problem. An ensemble is expected to fuse the group information and the specific information so to improve the generalization of the ECG beat annotator over both the global *RSVC* and the local ν *SVC*.

One can directly fuse the outputs of the two *SVMs* using some simple fixed rules, such as maximum rule (MAX), minimum rule (MIN) sum rule (SUM), product rule (PRD) [56, 70]. However, the outputs of *SVMs* are not posterior probabilities and appear to be in different ranges. Therefore, the outputs (confidence) of the two *SVMs* are not comparable directly due to the different training data used. One can estimate the posterior probabilities $P(y_j|x)$ by fitting sigmoid function [94], as introduced in Section (3.3.5)

Another option is to train the ensemble using a training set. Fusion rules such as stacking, decision template [56] can be used. But the selection of the training set has to be taken carefully. It is dangerous to use the same training set by the ensemble and the base classifiers [76]. Since the local ECG training set of ν *SVC* is more similar to the patient than that of the global training set of *BSVC*, a training set using the extra ECG data from the same patient is preferable. However, only normal ECG data can be used due to the requirement that no annotation of ECG beats is required for

the physician. Therefore, some artificial outlier data can be generated as described in section 4.4.1.4. But the number of the normal ECG data for training of the ensemble also has to be small as possible. So the training data set of the local νSVC can be added with some extra normal ECG data from the patient to form the training set of the ensemble. On one hand, the information from local ECG training data can be exploited further. On the other hand, overtraining of the ensemble can be avoided by incorporating extra normal ECG data from the patient into the training set of νSVC .

4.5.2 Experiments

4.5.2.1 Experimental Setting

The proposed *HKME*-based patient-adaptable ECG beat annotator was evaluated on MIT/BIH arrhythmia database. The experimental setting is similar to that of in section 4.4.2.1.

1. 22 recordings are selected as local training sets and test sets, in which the number of abnormal beats is significantly less than that of the normal beats, to be in agreement to the scenario in long-term monitoring of patients suffering from cardiovascular diseases. These recordings include # 100, 105, 106, 108, 114, 119, 121, 200, 203, 205, 208, 209, 210, 213, 215, 221, 222, 223, 228, 230, 233 and 234. Each of the 22 recordings is split into two sets.
 - The first 200 normal ECG beats in each of the 22 recordings (about 3 minutes) are used as the local training set to construct the νSVC s.
 - The first 350 normal ECG beats in each of the 22 recordings (about 5 minutes) are used as the training set to train the ensembles.
 - The second 1/2 of each of the 22 recordings (about 15 minutes or 1000 beats) is used as test set to evaluate the performance of the ECG annotators.
2. 10000 ECG beats (with half normal beats and half abnormal beats) from 22 recordings are used as global training set (DB_G) to train some classical binary

Table 4.6: Results (average $\pm$ standard deviation) of abnormal ECG beat annotation (all in percentage).

Classifiers	BCR	SEN	SPE	ACR
<i>RSVC</i>	80.3 ± 16.9	81.3 ± 24.5	79.3 ± 29.6	80.2 ± 26.1
ν <i>SVC</i>	83.6 ± 14.7	87.5 ± 22.5	79.7 ± 21.3	81.7 ± 18.5
<i>MAX</i>	86.2 ± 16.4	87.0 ± 21.9	85.4 ± 20.6	85.8 ± 19.5
<i>LDC</i>	86.5 ± 16.2	82.6 ± 27.3	90.5 ± 16.1	90.1 ± 14.1
<i>QDC</i>	83.1 ± 17.3	74.6 ± 35.1	91.6 ± 12.2	90.3 ± 10.3
<i>DET</i>	87.5 ± 15.0	83.3 ± 26.6	91.6 ± 13.7	91.2 ± 11.5
<i>ORA</i>	93.3 ± 10.6	93.2 ± 18.3	93.3 ± 12.9	93.7 ± 11.4

classifiers for comparison with the proposed *HKME*-based patient-adaptable ECG beat annotator. These recordings are # 101, 103, 109, 111, 112, 113, 115, 116, 117, 118, 122, 123, 124, 201, 202, 207, 212, 214, 220, 222, 231 and 232.

4.5.2.2 Results and Discussion

The annotation results of using the proposed *HKME* with different fusion rules, the global binary *SVM* and the local ν *SVC* are given in Table 4.6.

The reported results are averaged over 22 test ECG recordings. In accordance with AAMI recommendations to present the results by each patient, the annotation results of ν *SVC*, *RSVC* and the *HKME* (using DET fusion rule) in each recording are shown in Figure 4.12 and Table 4.8. Test sets #1 – #22 correspond to recording 100, 114, 119, 121, 200, 203, 205, 208, 209, 210, 213, 215, 221, 222, 223, 228, 230, 233, 234, 105, 106 and 108 in MIT/BIH arrhythmia database respectively.

Table 4.7: Number and percentage of test sets in which *HKME* outperforms both global *RSVC* and local ν *SVC* among all 22 test sets.

Fusion Rule	<i>MAX</i>	<i>LDC</i>	<i>QDC</i>	<i>DT</i>
in number	14	16	11	18
in percentage	63.6	72.7	50.0	81.8

- Overall Performance and Generalization

It is observed from Table 4.6, all the *HKME*s except using QDC rule outperforms both the global *RSVC* trained using the ECG data from a large patient-group and the local ν *SVC* trained using some normal ECG data from each patient. The generalization of both the global *RSVC* and the local ν *SVC*

Figure 4.12: Comparison of the annotation results of the global $RSVC$, local νSVC and $HKME$ (with DET fusion rule) in each recording of the test sets in terms of BCR.

can be improved when the information from these two sources are integrated properly by the $HKME$.

The best BCR achieved by $HKME$ is using DET rule, whose BCR is 7.2% and 3.9% higher than the global $RSVC$ and the local νSVC respectively. DET-based $HKME$ outperforms both SVM s in more than 80% of the test sets as reported in Table 4.7. The second best BCR achieved by $HKME$ is using LDC-based stacking rule, which outperforms both SVM s in about 72% of the test sets. The performance improvement in terms of BCR using MAX rule is observed in about 64% of the test sets. Its average of BCR is 5.9% and 2.6% greater than the global $RSVC$ and the local νSVC respectively. The only exception is QDC rule. This may be resulted by the fact that the covariance matrices for the classes are near singular sometimes, these QDC classifiers may fail when trying to estimate and invert the covariance matrices [56]. It is expected that its performance may be better if it is properly trained.

It can be observed that the performance of trained $HKME$ s such as DET and

Table 4.8: Annotation results of the *RSVC*, ν *SVC* and *HKME* with DET fusion rule in each recording of the 22 test sets.

Record #	<i>RSVC</i>				ν <i>SVC</i>				<i>HKME</i> (DET)			
	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN
100	3	19	0	1113	4	18	126	987	4	18	111	1002
105	79	23	199	1042	102	0	488	753	102	0	432	809
106	368	9	7	663	376	1	7	663	376	1	4	666
108	31	4	749	126	11	24	85	790	11	24	83	792
114	9	1	781	152	8	2	557	376	8	2	520	413
119	290	17	9	729	307	0	99	639	307	0	57	681
121	10	2	69	855	12	0	523	401	12	0	480	444
200	545	47	3	799	575	17	46	756	574	18	9	793
203	259	9	766	517	196	72	11	1272	193	75	10	1273
205	62	5	6	1261	67	0	152	1115	67	0	107	1160
208	557	73	13	875	619	11	353	535	608	22	126	762
209	17	108	2	1396	84	41	261	1137	20	105	2	1396
210	133	22	77	1108	154	1	183	1002	154	1	122	1063
213	187	97	0	1361	273	11	603	758	270	14	470	891
215	101	2	324	1271	103	0	200	1395	103	0	56	1539
221	172	2	6	1049	174	0	72	983	174	0	45	1010
222	197	138	469	497	298	37	690	276	281	54	589	377
223	362	11	548	398	348	25	112	834	346	27	92	854
228	178	8	8	874	155	31	5	877	156	30	4	878
230	64	7	69	1091	71	0	70	1090	71	0	47	1113
233	444	6	658	466	445	5	67	1057	445	5	25	1099
234	5	2	0	1373	7	0	51	1322	7	0	17	1356
Total	4073	612	4763	19016	4389	296	4761	19018	4289	396	3408	20371

LDC is better than that of the non-trained *HKME*, the MAX rule. It shows that the proper training of the fusion rules is helpful to the improvement of the ensemble over the base classifiers. These findings are similar to those in Chapter 3.

- Influence of the training set of the ensemble

An experiment was conducted to evaluate the influence of the size of the training set of the *HKME* ensemble to the performance of the ensemble. The BCR achieved by *HKME* with DET rule is illustrated in Figure 4.13, as the number of training data for the ensemble is varied from 50 to 450. The gray bars are those excluding the training data of the ν *SVC* and those black bars are those including the training data of the ν *SVC*. It is observed that the performance of

Figure 4.13: The influence of training set size to the performance of the *HKME* (DET) in terms of BCR.

the *HKME* is improved gradually with the increase of the number of training data first, but the performance improvement stops when the number of the training data is greater than 250 regardless whether including or excluding the training data of the νSVC . Therefore, the influence of the training data size of the ensemble to its performance is obvious. With more training data, higher performance is expected by the ensemble. However, more training data from each patient means more cost. Here the extra 150 normal ECG data seems a good tradeoff.

By including the training data of the local νSVC to the training set of the ensemble, the performance of the *HKME* ensemble is better than that of excluding the training data of the local νSVC , especially when the number of the extra normal ECG data is small (less than 150 here). It supports the claim in section 4.5.1.2 that the inclusion of the training data of νSVC in the training set of the ensemble is helpful to the performance of the *HKME*.

- Diversity and Performance of *HKME*

The relations of the improved BCR over *RSVC* and νSVC and the diversity

Figure 4.14: The relation between the improved BCR of the $HKMEs$ ($DETs$) over the global $RSVC$ (A) and local $\nu SVCs$ (B) and the plain disagreement measure.

measures are illustrated in Figure 4.14, Figure 4.15 and Figure 4.16.

- Plain Disagreement Measure (PDM): $PDM = \frac{N_{dis}}{N_{all}}$, where N_{dis} is the number of samples that two classifiers disagree and N_{all} is the number of all samples in the validation set. PDM varies from 0 to 1. The larger the value of P , the higher the diversity. This measure was recommended for ensemble feature selection in [88].

It is observed from Figure 4.14 that larger PDM (higher diversity) does not mean higher performance improvement of the ensemble. But higher performance improvement of the ensemble usually corresponds to larger PDM (higher diversity). It shows that high diversity is a necessary condition, but not a sufficient condition for the high performance improvement of $HKME$ ensemble.

- Double Fault Measure (DFM): Let N_{ij} be the number of patterns in the validation set for which the outputs of the two classifiers are i and j respectively.

$DFM = \frac{N_{00}}{N_{00}+N_{01}+N_{10}+N_{11}}$. It is the proportion of the patterns that have been misclassified by both classifiers. DFM varies from 0 to 1. The larger the value of DFM , the higher the diversity. This measure was recommended in [66].

Similar observation can be found in the relation between the Double Fault

Figure 4.15: The relation between the improved BCR of the $HKMEs$ ($DETs$) over the global $RSVC$ (A) and local $\nu SVCs$ (B) and the Double Fault measure (DFM).

Figure 4.16: The relation between the improved BCR of the $HKMEs$ ($DETs$) over the global $RSVC$ (A) and local $\nu SVCs$ (B) and the correlation coefficient measure (COR).

Measure (DFM) and the performance improvement of $HKME$ ensemble in Figure 4.15

- Correlation Coefficient Measure (COR): Let N_{ij} be the number of patterns in the validation set for which the outputs of the two classifiers are i and j respectively.

$$COR = \frac{N_{00}N_{11} - N_{01}N_{10}}{\sqrt{(N_{00} + N_{10})(N_{01} + N_{11})(N_{00} + N_{01})(N_{10} + N_{11})}}. \quad COR \text{ varies from } -1 \text{ to } 1.$$

The expectation of it is 0 for statistically independent classifiers. The value of COR is positive when the classifiers tend to classify the same patterns correctly, and it is negative when the classifiers tend to make errors on different patterns. This measure was used in [86].

It is observed that about half of the cases has positive correlation while the other half has negatively correlation. Though negatively correlated individual classifiers are recommended to improve the performance of ensembles in [146], it seems that most performance improvement is achieved by positively correlated classifiers in Figure 4.16. This may be due to the fact that only two classifiers (*BSVC* and ν *SVC*) are used in the *HKME* ensemble here and the two classifiers have to be trained to minimize the errors to precisely represent the information from two sources (patient-group and specific patient). Hence the two base classifiers tend to classify the same patterns correctly. Their disagreement implies the difference from the two information sources. These two base classifiers have to be not only different to each other (higher diversity), but also to be strong enough.

It is also observed that the correlation coefficients are usually small (within 20%). It shows that information from the patient-group and each specific patient are quite different to each other, which is the motivation of the *HKME* ensemble to address the problem of poor generalization of the ECG beat annotators.

- Computation complexity

The *HKME* is a combination of global *BSVC* and local ν *SVC*. Its computation complexity is the sum of those of the two *SVMs*. It seems that there is no problem in the computation of the *BSVC* and ν *SVC* from Table 4.5 if a desktop or a server is used in the long-term monitoring of the heart patients. If it is to be implemented in an embedded system or a PDA platform [147], some efforts may be necessary to further decrease the computation time. For example, a hierarchical structure may be employed in which a local ν *SVC* is used for most of the ECG beat annotation and the *HKME* is only used when the confidence of the local ν *SVC* is not high enough.

- Observation from Oracle

Note that an fusion rule called Oracle is reported in Table 4.6. Oracle (ORA) is the optimal case or an upper bound which an ensemble can reach rather

than a real ensemble. It assigns a correct class label to the pattern iff at least one individual *SVM* produces a correct class label [56]. Here it is used for comparison purpose only.

Compared with the best single *SVMs* in the *HKME* ensemble, the performance improvement of *DET* is 3.9% in terms of BCR. But the upper bound of performance improvement of the *HKME* ensemble (Oracle) is 9.7%. Hence there is still a 5.8% gap in terms of *BCR* between the *DET* ensemble and the oracle, which means there are still some space for further improvement of the *HKME* ensemble.

4.6 Concluding Remarks

In this chapter, two new approaches are proposed for patient-adaptable abnormal ECG beat annotation for long-term monitoring of patients suffering from cardiovascular diseases. These two patient-adaptable ECG beat annotators are developed to address the problem of poor generalization and imbalanced data problem.

One is the concept learning-based approach. A concept learning model, ν *SVC* is trained using only a small number of “normal” ECG beats from a patient, thus to learn the concept “normal beat” from these limited number of samples. The constructed ν *SVC* model can then be used to annotate “abnormal” ECG beats in long-term ECG recordings of the same patient. This approach can relieve physicians from annotating the ECG data beat by beat for training the patient-adaptable ECG annotator and is able to address the generalization problem in ECG signal annotation. Experimental results using 44 ECG recordings of MIT/BIH arrhythmia database show that the proposed concept learning model outperforms the other classifiers trained using the ECG data from a large patient-group. Furthermore, it is easy to implement and faster than the *RSVC* in the annotating the ECG beats from the patient.

The other is the hybrid kernel machine ensemble approach. The hybrid kernel machine ensemble consists of two types of kernel machines. One is the ν *SVC* trained using only some “normal” ECG beats from a specific patient to obtain specific reference information of the patient. The other is the *BSVC* trained using a large

database which consists of ECG beats from many patients to obtain standard reference information. Due to the different information represented in these two types of kernel machines, they perform differently when annotating the abnormal ECG beats from the specific patient. The integration of the two types of kernel machines in an ensemble is shown to perform better than using either of the kernel machines in annotating the abnormal ECG beats from the specific patient. Similar to the concept learning-based approach, this hybrid kernel machine ensemble can relieve the physicians from annotating the training ECG data beat by beat to train a local classifier and is able to improve the generalization. Experimental results using 44 ECG recordings of MIT/BIH arrhythmia database demonstrate the good performance of the proposed hybrid kernel machine ensemble and suggest its feasibility in practical clinical application.

Chapter 5

HKME-based Abnormal Region Detection in Colonoscopic Images

5.1 Introduction

In this chapter, an application of the HKME method to the detection of abnormal region in colonoscopic images is presented. Colonoscopy is a minimal invasive procedure of screening the colon and rectum using a colonoscope [148]. A typical colonoscope consists of a tiny CCD camera with a bright light fitted at the end of a long tube connecting to a viewing screen or television monitor at the other end, which displays a magnified image where the camera is pointing to (Figure 5.1). The procedure is used to look for signs of cancer in the colon and rectum and diagnose the causes of unexplained changes in the bowel such as inflamed tissue, abnormal growths, ulcers and bleeding etc. The procedure may take a long time and has to be done fully carefully so not to cause damage to the intestine. Analyzing colonoscopic images for clinical diagnosis of abnormalities relies on the experience and expertise of the medical experts, which need years of training to acquire. It is thus significant to develop a computer-assisted technique to help the screening process of these potentially lethal diseases by the healthcare provider. The intelligent instrumentation system based on such technique can facilitate medical experts to diagnose the abnormalities of patients in large numbers, particularly in screening procedures, and can also be used for training purpose.

Previous research on colonoscopic image analysis focused on the classification between normal tissues and tumors. However, few work has been done to discriminate normal tissues from different kinds of abnormalities including tumors in colonoscopic images, which is more significant for screening purpose. In fact, many categories of abnormalities can be seen in colonoscopic images, such as polyps, tumors, inflammation, bleeding, ulceration and diverticula etc.(Figure 5.2) and their image content show large variations. The abnormal regions usually do not occupy the whole image and vary in color, size and shape, which add more difficulties to the discrimination of the normal regions from the abnormal ones in colonoscopic images. In this chapter, a *HKME*-based approach using multi-size patches is proposed for detecting abnormal regions in colonoscopic images. Multiple sizes of patches provide multiple level visual cues of the image regions, which can help produce better perceptually agreeable segmentation. Represented as multi-size patches, the abnormal region detection in colonoscopic images turns into a binary classification problem to discriminate the patches from normal regions (normal class) and those from abnormal ones (abnormal class). Each pixel in a given image can be categorized as normal or abnormal using a trained patch-based classifier. Using multiple sizes of patches, multi-labels can be given to a pixel, the final label of the pixel can be obtained using the ensemble of these multiple classifiers based on different patch sizes.

The performance of the ensemble depends on the individual classifiers used. A

(a) Colonoscope [149]

(b) Colonoscopy system [150]

Figure 5.1: Colonoscope and colonoscopy system

Figure 5.2: A normal colonoscopic image and 5 colonoscopic images with different types of abnormalities.

set of individual classifiers have to be trained for the binary classification problem to discriminate the normal patches from those abnormal ones. This problem can be solved using a discriminative model, such as *BSVCs*. Such a *BSVC*-based abnormal region detection approach in colonoscopic images using multiple-size patches was published in [151] (A preliminary work by the author).

Rather than a typical binary classification problem, the abnormal region detection in colonoscopic images can also be treated as a concept learning problem. A lot of patterns from abnormal regions in colonoscopic images for each categories of abnormalities have to be collected for training a reliable classifier, which means the concept “abnormal” is not easy to learn. On the other hand, the normal patterns show smaller variations than those of the abnormal ones and are much easier to be obtained. This means the concept “normal” can be easier to learn. Therefore, the concept “normal” can be learned using a one-class classifier, such as ν *SVC* or *SVDD*. Such a concept learning-based abnormal region detection approach in colonoscopic images was published in [152] (Another preliminary work by the author). Trained using only the data from one class, ν *SVCs* try to find a decision boundary around the training data – called targets, which is different from the decision boundary of *BSVC* trained using the data from both normal and abnormal classes. As explained in Chapter 2, ν *SVC* tries to represent of target samples rather than for discrimination purpose. On one hand, multi-size patches produce multi-level cues of image content, which in turn produce a diverse feature set. On the other hand, the combination of the two different types of kernel machines ν *SVC* and *BSVC* can produce more diversity to the ensemble, which may further improve the abnormal detection in colonoscopic images. Experimental results show that the multi-size patch-based hybrid kernel machine ensemble method is superior to that of using single patch size

only for the abnormal region detection in colonoscopic images and can produce more perceptual agreeable image segmentation.

5.2 Related Work

5.2.1 Perceptual Image Segmentation

Detecting abnormal regions in colonoscopic images can be regarded as a perceptual image segmentation problem. Many methods have been developed for segmenting the medical images, such as thresholding, region growing and classification based methods etc. [153]. However, the partition of pixels in an image in most of these methods is usually based on low level cues, such as illumination, color, texture etc. and the resulted segmentation often disagree with the way that human-beings partition the image.

Recently, there are some attempts to segment the images by incorporating higher level knowledge. For example, image segmentation is regarded as a graph partitioning problem and a criterion called “Normalized Cuts” for segmenting the graph is proposed [154]. Some improved versions of “Normalized Cuts” have been developed, such as in [155], and it is used in medical images, such as Magnetic Resonance Images (MRI) [156]. However, it is very difficult to determine the number of segments in an image without any prior knowledge.

In [157], a binary classification model for segmentation is developed. Instead of using a single pixel for segmentation, they grouped pixels into “superpixels” which are roughly homogenous in shape and size. Then, these superpixels are grouped into different segments using a classifier. These superpixels provide a higher level cue for segmentation than single pixel and produce better segmentation results than that of using only single pixels.

In [158], the whole image (the scene context) was used as an extra source of (global) information. This is to help resolve local ambiguities for object detection and scene recognition.

Joic proposed an intermediate appearance and shape model between pixel and the whole image called “epitomes” [159], which is a miniature, condensed version of

an image containing the essence of the textural and shape properties of the image. They used square patches as epitomes.

The patch-based approach seems to be a good representation model for image segmentation in which the full image is cropped into a set of image patches and these patches can be classified into different categories corresponding to different types of segments. This approach has been used extensively in many applications, such as face detection [160, 161], object detection [162–165] and image segmentation [166], etc.

An open problem of patch-based approach is how to choose an appropriate patch size. Smaller patch size cannot capture sufficient information of the object and often leads to large detection error. Larger-size patches contain more information about the objects that match their size and achieve better detection, but fail to represent smaller objects. It is necessary to solve this problem to improve the object detection and perceptual image segmentation.

5.2.2 Colonoscopic Image Analysis

Previous research on colonoscopic image analysis focused on the classification between normal tissues and tumors. For examples, Krishnan *et al* proposed to characterize the colon based on some quantitative parameters which define the properties of the lumen of the colon [167]. These properties can be used to decide whether the colon lumen of interest belongs to an abnormal or normal category based on a fuzzy characterization rule base. In [168], curvature change of a haustra creases contour was used for detecting abnormalities in endoscopic images. This approach is useful for discriminating those abnormalities which can be characterized by the shape of the colon lumen, such as polyps and tumors. But it cannot be used for some other abnormalities which cannot be characterized by the shape of the colon lumen, such as bleeding. Wang *et al* proposed a lumen center detection approach and a lumen-oriented image segmentation algorithm for colonoscopic images [169]. Krishnan *et al* proposed to classify the colonoscopic images using neural networks based on some color histograms and spatial domain parameters and their experimental results based on 22 colonoscopic images indicated that self-organizing network produced good classification results [170]. They took whole colonoscopic image as a pattern for classification.

But this may not be appropriate because there are usually only some regions in a colonoscopic images showing abnormality while the other regions are normal. Zheng *et al* proposed to combine the information from color, texture and shape of the lumen for detect abnormalities in endoscopic images [171–173].

The classification between tumor lesions and normal tissue is usually implemented by neural networks. For example, Tjoa *et al* investigated back-propagation networks and the adaptive resonance theory networks in detecting abnormalities in colonoscopic images [174]. They employed multi-layer feed-forward neural network for colon status classification and proposed to improve computation efficiency using principal component analysis [175]. Wang *et al* employed a self-organizing map for the classification and segmentation of the colonoscopic images [176]. Karkanis *et al* employed a multilayer feed-forward neural network based on some statistical texture features for tumor detection in monochrome colonoscopic images [177]. Maroulis *et al* developed a detection system for colorectal lesions in endoscopy video-frames using multi-layer neural network as a classifier [178].

Magoulas *et al* proposed to improve neural networks using on-line learning to modified environmental conditions in colonoscopy images [179]. Furthermore, Magoulas *et al* presented an approach to detect tumors in colonoscopic images which is based on the synergy between unsupervised learning and neural networks [180]. Each pixel in the image was grouped into clusters using k-windows clustering and the data in each cluster are used to train a neural network. But it is difficult to determine a meaningful number of the clusters.

There are also some attempts to improve neural networks by using fuzzy logic. For examples, Kodogiannis developed a neural-fuzzy system based on texture features in 6 color channels for detecting abnormal lesions in endoscopic images captured by a swallowable imaging capsule [181]. Experimental results using 73 endoscopic images demonstrated its feasibility.

Most of the algorithms aforementioned try to detect the abnormal regions in colonoscopic images based on single pixel or full image. Karkanis *et al* investigated patch-based approach for polyp and tumor detection in colonoscopic images [177, 182–184]. Several patch sizes were tried separately and the one with the least detection error was selected to detect polyps and tumors in colonoscopic images. However, the

sizes of the polyps and tumors are different and the shapes of the tumors are often irregular (Figure 5.2). Therefore, using single-size patches may not be suitable for all types of abnormalities.

5.3 Proposed Methodology

Figure 5.3: The flowchart of *HKME*-based approach for abnormal region detection in colonoscopic images.

Figure 5.3 illustrates the flowchart of the proposed *HKME*-based approach using multi-size patches for abnormal region detection in colonoscopic images. The detail is as follows.

5.3.1 Image Region Representation Using Multiple-size Patches

As illustrated in Figure 5.2, the abnormal regions in colonoscopic images come from different categories and they vary in location, shape, color and size. The representation of these regions has to be considered carefully. It is similar to object detection in which the abnormal regions are the objects to be detected.

Figure 5.4: Patch-based image region representation.

Patch-based approach seems a good solution which turns the abnormal region segmentation into a binary classification problem [182]. As illustrated in Figure 5.4, each colonoscopic image can be cropped into a set of overlapping image patches and these image patches can be categorized as abnormal region class or normal region class by a classifier. The abnormal regions can thus be segmented from the normal ones.

An open problem in patch-based approach is what patch-size to choose. Compared to large ones, small-size patches (the extreme of a small-size patch is single pixel) can represent the image regions more precisely, but it contains less information of the image content than large-size patches (the extreme of large-size patch is the full image) and usually lead to larger classification error. The large-size patches contain more information of the object, but small abnormal regions in these patches may be missed. This is why it is very difficult to determine the appropriate patch-size to use.

In this chapter, a multi-size patch-based representation is proposed in which multi-size patches are used simultaneously to represent the image regions in colonoscopic images. Using patches of multiple sizes aims at overcoming the scale problem, i.e., an abnormal region may appear at different sizes in different images. Multi-size patches provide multiple-level representation of the image contents. At least some among all the patch sizes can better characterize the object. Hence, the integration of the detection result based on multi-size patches is expected to detect the abnormal regions more precisely than those based on single-size patches only. This is the novelty of this proposal.

5.3.2 Image Preprocessing

The original endoscopic images obtained by the endoscopy system are *RGB* images with the resolution of 256×256 pixels. Judge color similarity in *RGB* color space is difficult because it is not perceptually uniform space. Therefore, *CIELab* color space is used because it is a perceptual space where the difference can be calculated by Euclidian metrics.

CIELab colors are device-independent. The *Lab* space consists of three color channels. The first channel is Luminance (L), can range from 0 to a positive maximum number. A Lightness value of 0 equals black and a maximum value of equals white. So, the higher the value, the more vivid the color. The other two channels, a and b , represent color ranges. The channel a contains colors ranging from green to red and the channel b contains colors ranging from blue to yellow. This color space is defined with regard to the *CIEXYZ* color spaces, thus the *RGB* image is firstly transformed into *XYZ* color space by the following equations.

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.412453 & 0.357580 & 0.189423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (5.1)$$

Then the Lab color space can be obtained by

$$L = \begin{cases} 116(Y/Y_0)^{1/3} - 16, & Y/Y_0 > 0.008856 \\ 903.3Y/Y_0, & Y/Y_0 \leq 0.008856 \end{cases} \quad (5.2)$$

$$a = 500((X/X_0)^{1/3} - (Y/Y_0)^{1/3}) \quad (5.3)$$

$$b = 200((Y/Y_0)^{1/3} - (Z/Z_0)^{1/3}) \quad (5.4)$$

where the terms X_0, Y_0, Z_0 are the tristimulus values for the reference white [185].

The colonoscopic images in *RGB* color space are transformed into 3 bands in *CIELab* color space through which the color and luminance component can be processed and analyzed individually. The image is scanned across and cropped into a set of fixed sizes of patches respectively. The patches are overlapped by 50% to ensure

that no abnormal region is missed. Here 3 sizes of the image patches are investigated for abnormal region detection in the colonoscopic images, namely, 48×48 , 32×32 and 16×16 (pixels). Two sample of the image patches are illustrated in Figure 5.5. Feature can be then extracted from these image patches for classification.

Figure 5.5: Two examples of patches from colonoscopic images. The component L, a and b of the image patches are illustrated listed in a row.

5.3.3 Feature Extraction

Previous research have shown that color and texture features are useful for discriminating the abnormalities from the normal regions in colonoscopic images [173, 176, 177, 182]. Here both color and texture features are included in this research.

- **Color Features:** The color features are 2-dimensional(D) histograms of the components a and b in CIE_{Lab} color space. The number of bins of the histogram is 8 for 2-D histograms.
- **Textural Features:** Two-level Discrete Wavelet Transform (DWT)-based statistical features and 1-D histograms of luminance (the number of bins of the 1-D histogram is 16) are employed as textual features. The image patches are processed using two-level DWT. The mean and standard deviation of the absolute value of the approximate and detailed coefficients from the two-level DWT decomposition of the image patches in the 3 channels of the CIE_{Lab} color space are calculated as the textural features.

Altogether 128 features are extracted, giving rise to a feature vector of 128-D. Then the feature vectors from patches can be used to form the data set for classification.

A set X of N feature vector x_i , $X = \{x_i \in R^{128} | i = 1, 2, \dots, N\}$ for N patches, are labelled as $y_i \in \{+1, -1\}$ to indicate whether it is a normal patch or a patch containing abnormalities.

5.3.4 Learning *SVMs* for Image Patch Classification

5.3.4.1 *BSVC* Learning

It has been shown in Section 2.2 that *SVM* has good generalization by finding an optimal separating hyperplane which minimizes the classification errors made on the training set while maximize the “margin” between different classes. Given a training set X with N samples, the optimal separating hyperplane can be expressed as:

$$f(x) = \sum_{i=1}^N y_i \beta_i K(x_i, x) - b \quad (5.5)$$

where $K(\cdot)$ is a kernel function, b is a bias, β_i are the solutions of a quadratic programming problem to find the maximum margin. There are only a few training samples whose β_i s are non-zero. They are the Support Vectors, which are either on or near the separating hyperplane. The decision boundary, i.e. the separating hyperplane is along these support vectors, whose decision values $f(x)$ (equation 5.5) approach zero. Compared with the support vectors, the decision value of positive samples have larger positive values and those of negative samples have larger negative values. Therefore, the magnitude of the decision value can also be regarded as the confidence of the *SVM* classifier. The larger the magnitude of $f(x)$ is, the more confident the classification is.

Using overlapped image patches, each pixel in the patch can be classified as normal or abnormal by a *SVM* classifier corresponding to the patch size. Thus each pixel in the original image can have at least one label. If a pixel is classified differently by overlapped patches, the label of the patch that has the largest absolute decision value (confidence) is chosen as the label of that pixel.

5.3.4.2 ν SVC Learning

The classification between normal and abnormal patches can also be solved by a one-class classifier, such as ν SVC. A ν SVC can be trained using the data from normal image patches for each patch size. Using overlapped image patches, each pixel in the patch can be classified as normal or abnormal by the trained ν SVC classifier corresponding to the patch size. Thus each pixel in the original image can have at least one label. If a pixel is classified differently by overlapped patches, the label of the patch that has the largest confidence is chosen as the label of that pixel.

5.3.4.3 Decision Fusion Using HKME

Since there are many kinds of abnormalities in colonoscopic images showing large variation, many patterns from abnormal regions in colonoscopic images have to be collected for training a reliable classifier and it is difficult to collect. This leads to an imbalanced data problem. One class – “normal” has many training samples and is easier to model, while the other class – “abnormal” is difficult to model because it has more diverse distributions than the normal class. Therefore, ν SVC is very suitable for this problem. As a recognition-based model, ν SVC tries to describe the target data rather than for discrimination purpose, it can handle the problem of missing information. However, ν SVC is often inferior to BSVC for discrimination purpose. There is a need to combine these two types of kernel machines for this problem.

A set of 2-SVCs can be constructed for the classification, while ν -SVCs can be used to provide further decision information. The classification results of the two kernel machines can be aggregated using an ensemble. The different natures of the two types of SVMs adds more diversity to the ensemble, which may further improve the performance of the ensemble. The fusion rules of the SVMs include maximum, average, product, decision template, stacking and majority voting etc, which can be found in section 3.3.4.

5.4 Experimental Results and Discussions

5.4.1 Data Preparation

The proposed approaches were evaluated using a database which consists of 58 clinically obtained colonoscopic images. There are 12 normal images and 46 images with abnormal regions. The abnormal regions mostly occupy only some parts of the whole image and the abnormalities include polyps, tumors, inflammation, bleeding, ulceration and diverticula etc. Figure 5.2 shows a normal colonoscopic image and 5 colonoscopic images with different types of abnormalities. The images are RGB images with the resolution of 256×256 pixels. The images were processed following the procedure in Section 5.3. The pixels in the original images were manually labelled to provide the ground truths. The detection results were compared with the ground truth and evaluated using the following criteria.

5.4.2 Evaluation Measure

The evaluation criteria are specificity (SPE), sensitivity (SEN) and Balanced classification rate (BCR). Where SPE is the fraction of normal regions detected among all the normal regions, SEN is the abnormal regions detected among all the abnormal regions and BCR is the weighted average of SPE and SEN .

$$BCR = \lambda SPE + (1 - \lambda)SEN \quad (5.6)$$

where $\lambda \in [0, 1]$ can be tuned to favor SPE or SEN . Smaller λ favors more on SEN , which means that the error on the abnormal class is punished more seriously. On the contrary, larger λ favors more on SPE , which means that the error on the normal class is taken more seriously. At the extreme case, only SEN or SPE will be considered when λ is 0 or 1 respectively. $\lambda = 0.5$ is used here, so that SPE and SEN are treated as equally important. Other values might be selected with respect to the requirement of the medical experts.

Table 5.1: Results of image patch classification.

Patch Size	16×16	32×32	48×48
<i>SVM</i>	76.3%	81.7%	86.1%
MFNN	74.4%	79.1%	82.9%

5.4.3 Classification Based on Single Patch Size

The performance of using *BSVM* with Gaussian RBF kernel is compared to that of using Multi-layer Feed-forward Neural Network (MFNN) [175, 178] for discrimination between abnormal and normal patches in colonoscopic images. In the experiment, 58 colonoscopic images are used. The numbers of collected image patches of size 48×48 , 32×32 and 16×16 are 2000, 1998 and 2006 respectively. The pixels in the original image are manually labelled to provide the ground truths. The patch containing mostly abnormal region is labelled as a positive sample, otherwise, it is labelled as a negative sample. A leave-one-out experiment was performed for the classification. Each time, the image patches from one of the original colonoscopic images were selected as a test set and the patches from the other 57 images were used for training. The experiment was performed 58 times, and the average of the accuracy rates of total 58 results was taken as the final accuracy rate of the classification of patches. The results are listed in Table 5.1.

Obviously, *SVM* achieved higher classification accuracy rate than MFNN. In addition, the leave-one-out experiment is repeated 5 times for 16×16 patches, the average accuracy rate of MFNN is $73.5 \pm 7.4\%$, while that of *SVM* is $76.3 \pm 0\%$. It shows that the stability of *SVM* is also better than MFNN. Therefore, *SVM* is justified to be used as the basic classifier for the abnormal region detection in colonoscopic images.

5.4.4 Detecting Abnormal Region Using Learned *HKME*

In this experiment, 46 colonoscopic images with multiple categories of abnormal regions and 12 normal ones are used. The numbers of collected image patches for training of 48×48 , 32×32 and 16×16 (pixels) patches are 2002, 2090 and 2126 respectively. The pixels in the original image are manually labelled as the ground truth for comparison. The patches containing mostly abnormal region were labelled as a positive sample, otherwise, a negative one. A leave-one-out experiment was

performed to evaluate the performance of the proposed method for abnormal region detection in colonoscopic images. In each round, one of the colonoscopic images was selected for testing and the patches from other 57 images were used for training. The experiment was repeated 58 times, the detected results were compared to the ground truth image and the average value of the total 58 results was taken as the final result.

The detection results of 4 colonoscopic images are illustrated in Figure 5.6 and 5.7. Note that the result of an ensemble called oracle is also listed, which assigns a correct label to the pattern if any of the single *SVMs* assigns a correct label [56]. In fact, it is an upper bound which an ensemble can reach. It can shed some light on the diversity of the ensembles.

5.4.4.1 ν SVC vs BSVC Using Single Patch Size

Table 5.2: Results of abnormal region detection using single patch sizes.

Patch size	Classifier	<i>BCR</i>	<i>SPE</i>	<i>SEN</i>
48×48	2-SVC	0.744	0.675	0.813
48×48	ν -SVC	0.539	0.991	0.088
32×32	2-SVC	0.738	0.675	0.802
32×32	ν -SVC	0.546	0.998	0.094
16×16	2-SVC	0.745	0.668	0.822
16×16	ν -SVC	0.538	0.946	0.094

In Table 5.2, it is observed that *BSVCs* outperform ν *SVC* in all the cases which agrees with the postulate that discriminative models are superior to that of recognition-based models. *BSVCs* achieved *BCR* around 74%, while ν *SVCs* achieved only 55%. The ν *SVCs* have a very high *SPE*, but almost completely fail for *SEN*. This may be resulted that the training set size used for ν *SVC* was too small and it also suffered from the curse of dimensionality. Compared to ν *SVCs*, *BSVC* have higher *SEN* while much less *SPE*, which may be good for add more diversity to the ensembles. The best *BCR* is 74.5% which was achieved using patches of size 16×16 .

5.4.4.2 Multi-size Patch Ensemble of ν SVCs or BSVCs

Table 5.3 illustrates the detection results of 3 patch size ensembles using ν *SVCs* or *BSVCs* separately. Obviously, the best ensembles outperforms that of the best

Table 5.3: Detection results (in terms of BCR) of abnormal region detection using different patch sizes and ensemble schemes. The ensembles are constructed using same types of SVM s, $BSVC$ or νSVC .

Ensemble	MAX	AVG	PROD	MV	DET	LDC	QDC	Oracle
$BSVC$	0.737	0.751	0.745	0.751	0.753	0.763	0.765	0.868
νSVC	0.532	0.551	0.535	0.551	0.538	0.533	0.536	0.569

SVM s using single patch size, which supports the claim that multi-size patch-based SVM ensemble can achieve better abnormal region detection in colonoscopic images. Due to the poor performance of individual νSVC s, the improvement of their ensemble is limited although there are still some.

5.4.4.3 $HKME$ Using Single-size Patches

Table 5.4: Detection results (in terms of BCR) of abnormal region detection using same patch sizes by $HKME$.

Patch Size	MAX	AVG	PROD	MV	DET	LDC	QDC	Oracle
48×48	0.551	0.551	0.551	-	0.744	0.746	0.540	0.903
32×32	0.556	0.556	0.556	-	0.738	0.741	0.548	0.900
16×16	0.563	0.563	0.563	-	0.745	0.745	0.736	0.899

Table 5.4 shows the detection results of the ensemble of a νSVC and a $BSVC$ based on single-size patches. Only DET and LDC achieved BCR comparable to the best single classifier and the performance of other ensembles did not outperform the best single classifier. This may be due to the fact that the νSVC and $BSVC$ are trained using the same features, which limit the performance of this scheme.

5.4.4.4 $HKME$ Using Multi-size Patches

Table 5.5 illustrates the detection results of the $HKME$ ensemble of $BSVC$ s using all 3 patch sizes plus 1 to 3 νSVC (s) trained using 1 to 3 patch size(s). Most of the ensembles shows improvement over the best single SVM based on single-size patches. The performance of LDC and AVG outperforms others. Figure 5.6 and 5.7 illustrates the result of the ensemble of $BSVC$ s using all 3 patch sizes and a νSVC (s) trained using patches with size of 48×48 . Obviously, the detection results by the $HKME$ ensemble is closer to the ground truth compared to those using single-size patches.

Table 5.5: Detection results (in terms of BCR) of abnormal region detection using different patch sizes and $HKME$. Row $A + *$ are the results of ensembles using all 3-size patches learned by $BSVC$ s plus 1-size or 2-size patches learned by νSVC (s) (1 for 48×48 , 2 for 32×32 , 3 for 16×16). Row ALL are the ensemble results using all 3-size patches and both $BSVC$ and νSVC .

Ensemble	MAX	AVG	PROD	MV	DET	LDC	QDC	Oracle
A+1	0.705	0.761	0.768	0.751	0.753	0.765	0.766	0.953
A+2	0.541	0.761	0.659	0.751	0.753	0.764	0.667	0.953
A+3	0.588	0.754	0.730	0.751	0.753	0.756	0.730	0.953
A+1+2	0.539	0.769	0.598	0.765	0.753	0.765	0.542	0.953
A+1+3	0.587	0.765	0.656	0.765	0.753	0.763	0.743	0.953
A+2+3	0.538	0.762	0.572	0.765	0.753	0.763	0.559	0.953
ALL	0.537	0.565	0.548	0.704	0.751	0.764	0.549	0.955

5.4.4.5 Observation from the Oracle

As aforementioned, the BCR produced by Oracle can be regarded as a diversity measure to some extent. The larger the value of BCR of Oracle, the more diverse of the ensemble. The BCR achieved by Oracle is listed in the last column of Table 5.3, 5.4 and 5.5. It is observed that the BCR of Oracle is only 86.8% for $BSVC$ ensemble and 56.9% for νSVC ensemble trained using multi-size patches respectively. The BCR of the Oracle of using one patch size only and combining $BSVC$ and νSVC is increased to about 90%, which shows that it adds some diversity but the increase is not very significant. The detection results of Oracle 2 in Figure 5.6 is using the ensembles of three $BSVC$ s using all 3 patch sizes and a νSVC using 48×48 patches and those of Oracle 1 in Figure 5.6 is using the ensembles of three $BSVC$ s trained by all 3 patch sizes. Obviously, the results of Oracle 2 is more similar to the ground truth than those of oracle 1. Therefore, the $HKME$ s of $BSVC$ s using all 3 patch sizes plus one or more νSVC (s) significantly improved the diversity, which increased the BCR of Oracle to more than 95%. It supports the claim that the multi-size patch-based $HKME$ ensemble produces higher diversity.

Of course, there is still some difference between the performance of $HKME$ and that of the *Oracle*. There is still scope for the ensemble to improve.

Figure 5.6: Detection results of 4 colonoscopic images. The regions in white are normal regions detected and the regions in black are abnormal ones.

5.5 Concluding Remarks

In order to deal with the variation in location and size of the abnormal regions in colonoscopic images, a multiple-size patch-based representation is proposed in this

Figure 5.7: Detection results of 4 colonoscopic images. The regions in white are normal regions detected and the regions in black are abnormal ones.

chapter. Larger-size patches contain more information about the abnormal region which matches its size and smaller-size patches are good for represent small abnormal regions. At least some among all the patch-size can better characterize the abnormal

regions when multiple-size patches are used simultaneously.

A new *HKME*-based method is proposed for detecting abnormal regions in colonoscopic images using multi-size patch-based representation. Represented as multi-size patches, the abnormal region detection in colonoscopic images is formulated as a binary classification problem to discriminate the patches from normal regions and those from abnormal ones. Each pixel in the given images can be categorized as normal or abnormal using a trained patch-based *BSVC* classifier. Using multi-size of patches, multi-labels can be given to each pixel, the final label of the pixel can be obtained using the ensemble of these multiple classifiers based on different patch size. Due to the use of multiple sizes of image patches, the proposed method can produce more perceptually agreeable detection results.

Furthermore, it is observed that an imbalanced data problem exists in colonoscopic images in which the area of normal regions is much larger than that of the abnormal ones. Therefore, recognition-based one-class *SVMs* are trained to combine with the discriminative binary *SVMs* in an ensemble to improve the abnormal detection in colonoscopic images. Exploiting the multi-size patch-based image region representation and complementary nature of the two types of *SVMs*, the proposed ensemble can produce better detection results than that of using single-size patch only. *HKME* shows marginal improvement over the binary *SVM* ensemble for the abnormal region detection. This may be due to the inferior performance of the one-class *SVM*. It is very hard to combine two classifiers whose performances are significantly different to each other [186]. This is known as the imbalanced classifier problem. More discriminative features need to be investigated to improve the classification by one-class *SVM*, so to improve the *HKME* as well. Among the investigated fusion rules, *LDC*-based stacking performs the best. Experimental results show the good performance improvement of the proposed ensemble and thus support its feasibility in clinical application.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

A kind of imbalanced data problem exists in computer-assisted diagnosis, in which the majority class compactly clustered while the minority class scattered in the input space. A new learning algorithms, *HKME* is proposed to address this problem by combining a discriminative *BSVC* and a recognition-based ν *SVC* in an ensemble. The algorithm is applied to two medical applications, one is annotating abnormal ECG beats for long-term monitoring heart patients and the other is detecting abnormal regions in colonoscopic images. The proposed *HKME* can be further extended to other applications such as image retrieval and object detection etc. Several related new methods are developed and summarized as follows.

1. A new learning algorithm, *HKME* is proposed which can be used to address the imbalanced data problem aforementioned. This algorithms is designed to combine a discriminative *BSVC* and a recognition-based ν *SVC* in an ensemble. On one hand, a *BSVC* is trained using the data from both majority and minority class. So it benefits from the information from both classes. But it suffers from the adverse effect due to the poorly represented minority class. On the other hand, a ν *SVC* is trained using the the data from majority class only. So it avoids the adverse effect of poor representation of the minority class. However, it is not highly discriminative since the data from the minority class is left totally unused. Exploiting the complementary properties of these two

types of kernel machines, the integration of both kernel machines using an ensemble is shown to outperform using either of them respectively. *HKME* is a combination of one-class classifier and two class classifier and its property is in between the two kinds of classifiers which is thus regarded as one-and-half classifier. The proposed *HKME* is evaluated using some artificial and real data sets. Experimental results show that *HKME* can outperform both ν *SVC* and *BSVC* when the two classes are imbalanced and the assumption of *HKME* is satisfied.

2. A patient-adaptable ECG beat annotator is developed using concept learning model for long-term monitoring of heart patients. The proposed approach aims to fill in the gap between the patient group and single patient. A concept learning model- ν *SVC* or *SVDD* can be constructed to learn the concept “normal” beats using only some “normal” ECG beats from a patient. The learned model can then be used to annotate the ECG beats from the same patient. Experimental results using MIT/BIH arrhythmia database show the proposed approach is good in generalization, easy to implement and quick in computation.
3. A *HKME*-based patient-adaptable ECG beat annotator is proposed for long-term monitoring of heart patients. This is motivated by the fact that physicians consider both the standard reference value of the patient group and the specific reference value of each patient in examining the ECG recordings of the patient. A ν *SVC* can be used to learn the specific reference value from each patient aforementioned. At the same time, a *BSVC* can be trained to learn the standard reference value from the ECG data of a patient group. The combination of these two sources using *HKME* is shown to further improve the classification. Experimental results show the good performance of the proposed approach and support its potential in real clinical application.
4. A patch-based *BSVC* ensemble is proposed to detect abnormal regions in colonoscopic images using multi-size patches. A multi-size patch-based representation is developed in order to address the scale and size problem in the patch-based approach. It is analogous to object detection in which the ob-

jects (abnormal regions) vary in size and location across the images. Multi-size patches are used simultaneously to represent the objects. At least some among all the patch size can better characterize the object. A discriminative *BSVC* is trained for each patch size. Multiple decisions can be given to each pixel in the image using multi-size patches. The final label is determined by the ensemble of the *BSVCs*. Experimental results show that the proposed approach is able to produce more perceptually agreeable detection in colonoscopic images.

5. A *HKME*-based approach is developed for abnormal region detection in colonoscopic images using multi-size patches. The objective is to address the imbalanced data problem in colonoscopic image analysis in which the normal regions dominant and similar to each other while the abnormal regions are far less than the normal ones and show larger variance. The proposed approach can produce better detection results than that of using single patch size as supported by the experimental results.

6.2 Recommendations for Further Research

Ensemble of kernel machines for imbalanced data problem in biomedical applications and other related problems still faces many challenges, which need to be explored further in the future. Some recommendations for further research are summarized as follows.

- The fusion rule for combining the two types of kernel machines in *HKME* needs to be further investigated. New approaches may be needed to improve the performance of the ensemble.
- *HKME* is using the estimated posterior probabilities of the *SVMs*. More precise estimation methods is needed to be developed to ease the integration of the *SVMs*.
- Good classification is based on efficient features. New feature definition and feature extraction need to be investigated to improve the classification in colonoscopic image analysis and ECG beat detection.

- It is promising to apply *HKME* to the other applications such as object detection and image retrieval to solve similar imbalanced data problem in those areas.

To sum up, research effort has been taken on developing new learning algorithms and implementing the algorithms to solve challenging biomedical signal / image processing problems, particularly for colonoscopic image analysis and ECG beat detection. It is believed that significant contributions have been made in the present research work, paving the way for intelligent diagnosis and screening for abnormalities in colonoscopic images and ECG signals and rendering efficient and cost-effective services in healthcare delivery in the near future.

Author's Publications

1. Peng Li, Shankar M. Krishnan, Kap Luk Chan and Yan Gao, "Abnormal Region Detection in Colonoscopic Images using Novelty Detection Technique", 7th International Workshop on Advanced Image Technology (IWAIT 2004), Singapore, 12-13, January, 2004.
2. Peng Li, Kap Luk Chan, Shankar M. Krishnan and Yan Gao, "Detecting Abnormal Regions in Colonoscopic Images by Patch-based Classifier Ensemble", 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, United Kingdom, 23rd-26th, August, 2004, pp:774-777.
3. Peng Li, Kap Luk Chan and Y. W. Chan, "A Mixed SVM-Based Hierarchical Learning Approach for Abnormal ECG Beat Recognition", 1st International BioEngineering Conference (IBEC 2004), Singapore, 8th-10th, September, 2004.
4. Peng Li, Kap Luk Chan, Sheng Fu and Shankar M. Krishnan, "An Abnormal ECG Beat Detector Approach for Long-term Monitoring of Heart Patients based on Hybrid Kernel Machine Ensemble", International Workshop on Multiple Classifier Systems (MCS 2005), Monterey Bay, Seaside, California, USA, 13-15, June, 2005, pp:346-355.
5. Peng Li, Kap Luk Chan and Shankar M. Krishnan, "Learning A Multi-size Patch-based Hybrid Kernel Machine Ensemble for Abnormal Region Detection in Colonoscopic Images", IEEE Computer Society International Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20-26, June, 2005, pp:670-675.
6. Peng Li, Hai Rong Sun, Kap Luk Chan, Li Hui Zhou and Shankar M. Krishnan, "Stacking of SVMs for Abnormal Region Screening in Endoscopic Images Using Multi-size Patches", The Fourth International Conference on Machine Learning and Cybernetics (ICMLC 2005), Guangzhou, China, 19-21 August, 2005.
7. Peng Li, Kap Luk Chan, Sheng Fu and Shankar M. Krishnan, "A Concept Learning-based Patient-adaptable Abnormal ECG Beat Detector for Long-term

Monitoring of Heart Patients”, accepted as a book chapter in Neural Networks in Healthcare: Potentials and Challenges, IDEA Group Inc. 2005 (in press).

8. Peng Li, Kap Luk Chan and Wen Fang, “Hybrid Kernel Machine Ensemble for Imbalanced Data Sets”, 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, China, 20-24 August, 2006 (accepted).

Bibliography

- [1] L. I. Kuncheva, “Combining classifiers: Soft computing solutions,” in *Lecture Notes in Pattern Recognition*, S. Pal and A. Pal, Eds., pp. 427–452, 2001.
- [2] V. Vapnik, *The Nature of Statistical Learning Theory*, New York, 1995.
- [3] N. Japkowicz, C. Myers, and M. Gluck, “A novelty detection approach to classification,” ” in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 518–523, 1995.
- [4] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
- [5] D. Tax and R. Duin, “Support vector data description,” *Pattern Recognition Letters*, vol. 20, no. 11-13, pp. 1191–1199, December 1999.
- [6] D. Tax, “One-class classification: concept-learning in the absence of counter-examples,” ASCI Dissertation Series, Delft University of Technology, June 2001.
- [7] B. Raskutti and A. Kowalczyk, “Extreme re-balancing for SVMs: a case study,” *SIGKDD Explorations*, vol. 6, no. 1, pp. 60–69, 2004.
- [8] N. Japkowicz, “The class imbalance problem: Significance and strategies,” ” in *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI’2000)*, pp. 111–117, 2000.
- [9] R. Mark and G. Moody, “MIT-BIH arrhythmia ECG database,” 1997. [Online]. Available: <http://ecg.mit.edu/>

- [10] B. Schölkopf and A. J. Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*, 2002.
- [11] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, 2000.
- [12] K.-R. Müller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf, “An introduction to kernel-based learning algorithms,” *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, May 2001.
- [13] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, “Kernel pca and de-noising in feature spaces,” ” in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pp. 536–542, 1999.
- [14] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, “Fisher discriminant analysis with kernels,” in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds., pp. 41–48, 1999.
- [15] F. Bach and M. Jordan, “Kernel independent component analysis,” *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 1–48, January 2003.
- [16] J. Peng, D. Heisterkamp, and H. Dai, “Adaptive quasiconformal kernel nearest neighbor classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 656–661, May 2004.
- [17] S. B. Gokturk, C. Tomasi, B. Acar, C. F. Beaulieu, D. Paik, B. J. Jeffrey, J. Yee, and S. Napel, “A statistical 3D pattern processing method for computer aided detection of polyps in CT colonography,” *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1251–1260, December 2001.
- [18] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, “A support vector machine approach for detection of microcalcifications,” *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1552–1563, Dec. 2002.

- [19] S. Osowski, L. Hoai, and T. Markiewicz, "Support vector machine-based expert system for reliable heartbeat recognition," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 582–589, April 2004.
- [20] D. Strauss, G. Steidl, and J. Jung, "Arrhythmia detection using signal-adapted wavelet preprocessing for support vector machines," *Computers in Cardiology*, pp. 497–500, 2001.
- [21] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–450, November 2002.
- [22] V. Vapnik, *Statistical Learning Theory*, New York, 1998.
- [23] C. M. Bishop, *Neural networks for pattern recognition*, Oxford, U.K., 1995.
- [24] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," " in *Advances in kernel methods: support vector learning*, pp. 185–208, 1999.
- [25] "<http://www.kernel-machines.org/>," A collection of literature, software and web pointers about SVM and Gaussian processes.
- [26] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," " in *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, pp. 481–492, 1951.
- [27] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.
- [28] M. Kubat, R. Holte, and S. Matwin, "Detection of oil-spills in radar images of sea surface," *Machine Learning*, vol. 30, pp. 195–215, 1998.
- [29] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalances data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, February 2004.

- [30] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: One sided selection,” ” in *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186, 1997.
- [31] A. Kolcz, A. Chowdhury, and J. Alsepector, “Data duplication: an imbalance problem?” in *The Twentieth International Conference on Machine Learning (ICML-2003) Workshop on Learning from Imbalanced Data Sets II*, 2003.
- [32] A. S. Nickerson, N. Japkowicz, and E. Milios, “Using unsupervised learning to guide resampling in imbalanced data sets,” ” in *Eighth International Workshop on Artificial Intelligence and Statistics*, January 2001.
- [33] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Artificial Intelligence Research*, no. 16, pp. 321–357, 2002.
- [34] C. Drummond and R. C. Holte, “C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling,” ” in *Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, July 2003.
- [35] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, “New applications of ensembles of classifiers,” *Pattern Analysis and Applications*, vol. 6, no. 3, pp. 245–256, 2003.
- [36] G. Weiss and F. Provost, “The effect of class distribution on classifier learning,” Department of Computer Science, Rutgers University, 2001. [Online]. Available: <http://citeseer.ist.psu.edu/article/weiss01effect.html>
- [37] A. Estabrooks, T. Jo, and N. Japkowicz, “A multiple resampling method for learning from imbalanced data sets,” *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.
- [38] G. J. Karakoulas and J. Shawe-Taylor, “Optimizing classifiers for imbalanced training sets,” ” in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pp. 253–259, 1999.

- [39] K. Veropoulos, N. Cristianini, and C. Campbell, "Controlling the sensitivity of support vector machines," " in *Proc. of the International Joint Conference on Artificial Intelligence, (IJCAI99), Stockholm, Sweden, 1999*. [Online]. Available: <http://citeseer.nj.nec.com/veropoulos99controlling.html>
- [40] H. Shin and S. Cho, "How to deal with large dataset, class imbalance and binary output in SVM based response model," " in *the Korean Data Mining Conference*, pp. 93–107, Dec. 2003.
- [41] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," " in *The Twentieth ICML Workshop on Learning from Imbalanced Datasets*, pp. 49–56, August 2003.
- [42] N. A. Murshed, F. Bortolozzi, and R. Sabourin, "A fuzzy ARTMAP-based classification system for detecting cancerous cells, based on the one-class problem approach," " in *Proceedings of the 13th International Conference on Pattern Recognition 1996*, vol. 3, pp. 478–482, Aug 1996.
- [43] A. Hojjatoleslami, L. Sardo, and J. Kittler, "An RBF based classifier for detection of microcalcifications in mammograms with outlier rejection capability," " in *International Conference on Neural Networks*, vol. 3, pp. 1379–1384, 1997.
- [44] P. Hayton, B. Scholkopf, L. Tarassenko, and P. Anuzis, *Support vector novelty detection applied to jet engine vibration spectra. Advances in Neural Information Processing Systems*, ch. 13, pp. 946–952, 2001.
- [45] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *Journal of Machine Learning*, no. 2, 2001.
- [46] D. Tax and R. Duin, "Image database retrieval with support vector data description," " in *Proceedings of the Sixth Annual Conference of the Advanced School for Computing and Imaging, ASCI Delft*, June 2000.
- [47] D. Xin, Z. Wu, and W. Zhang, "Support vector domain description for speaker recognition," " in *Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pp. 481–488, 2001.

- [48] D. Xin and Z. Wi, "Speaker recognition using continuous density support vector machines," *Electronics Letters*, vol. 37, no. 17, pp. 1099–1101, Aug 2001.
- [49] M. Markou and S. Singh, "Novelty detection: a review-part 1: statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.
- [50] —, "Novelty detection: a review-part 2: neural network based approaches," *Signal Processing*, vol. 83, no. 12, pp. 2499–2521, 2003.
- [51] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., 2001.
- [52] M. Gal-Or, J. H. May, and W. E. Spangler, "Assessing the predictive accuracy of diversity measures with domain-dependent asymmetric misclassification costs," *Information Fusion Journal (Special issue on Diversity in Multiple Classifier Systems)*, vol. 6, no. 1, p. 3748, 2005.
- [53] J. A. SWETS, "Measuring the accuracy of diagnostic systems," *Science*, vol. 24, no. 48, pp. 1285–1293, 1988.
- [54] L. J. Cao, H. P. Lee, and W. K. Chong, "Modified support vector novelty detector using training data with outliers," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2479–2487, 2003.
- [55] X. Lu, N. Zheng, and S. Zheng, "Support vector machines for information retrieval," *JOURNAL OF XI AN JIAOTONG UNIVERSITY*, vol. 37, no. 6, pp. 581–584, June 2003.
- [56] L. I. Kuncheva, J. Bezdek, and R. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.
- [57] K. Woods, P. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 405–410, APR 1997.
- [58] T. G. Dietterich, "Ensemble methods in machine learning," " in *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pp. 1–15, 2000.

- [59] R. E. Schapire and Y. Singer, “Improved boosting algorithms using confidence-rated predictions,” *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.
- [60] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Y. Bang, “Pattern classification using support vector machine ensemble,” ” in *16th International Conference on Pattern Recognition*, vol. 2, pp. 160–163, 2002.
- [61] K. Tumer and N. C. Oza, “Input decimated ensembles,” *Pattern Analysis and Applications*, vol. 6, no. 1, pp. 65–77, 2003.
- [62] H. Hao, C.-L. Liu, and H. Sako, “Confidence evaluation for combining diverse classifiers,” ” in *Seventh International Conference on Document Analysis and Recognition*, pp. 760–764, 2003.
- [63] R. Bryll, R. Gutierrez-Osuna, and F. Quek, “Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets,” *Pattern Recognition*, vol. 36, pp. 1291–1302, 2003.
- [64] L. Kuncheva and L. C. Jain, “Designing classifier fusion systems by genetic algorithms.” *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 4, pp. 327–336, 2000.
- [65] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S. Chen, “Lung cancer cell identification based on artificial neural network ensembles,” *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25–36, 2002.
- [66] G. Giacinto and F. Roli, “Design of effective neural network ensembles for image classification processes,” *Image Vision and Computing Journal*, vol. 19, no. 9-10, pp. 699–707, August 2001.
- [67] C. Y. Suen and L. Lam, “Multiple classifier combination methodologies for different output levels,” ” in *MCS ’00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pp. 52–66, 2000.
- [68] T. Windeatt and R. Ghaderi, “Binary labelling and decision-level fusion.” *Information Fusion*, vol. 2, no. 2, pp. 103–112, 2001.

- [69] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Special issue of the journal of INFORMATION FUSION on Diversity in Multiple Classifier Systems*, vol. 6, pp. 63–81, 2005.
- [70] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.
- [71] R. P. W. Duin and D. M. J. Tax, "Experiments with classifier combining rules," in *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pp. 16–29, 2000.
- [72] S. B. Cho and J. H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 25, no. 2, pp. 380–384, 1995.
- [73] Y. Lu and F. Yamaoka, "Fuzzy integration of classification results," *Pattern Recognition*, vol. 30, no. 11, pp. 1877–1891, 1997.
- [74] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, pp. 79–87, 1991.
- [75] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.
- [76] R. P. W. Duin, "The combining classifier: To train or not to train?" in *Proceedings of 16th International Conference on Pattern Recognition*, vol. 2, pp. 765–770, June 2002.
- [77] F. Roli, J. Kittler, G. Fumera, and D. Muntoni, "An experimental comparison of classifier fusion rules for multimodal personal identity verification systems," in *MCS '02: Proceedings of the Third International Workshop on Multiple Classifier Systems*, pp. 325–336, 2002.
- [78] F. Roli and G. Giacinto, "Design of multiple classifier systems," in *Hybrid Methods in Pattern Recognition*, A. H. Bunke, Ed., 2002.

- [79] L. Breiman, “Bias, variance, and arcing classifiers,” Statistics Department, University of California, 1996. [Online]. Available: <http://citeseer.ist.psu.edu/breiman96bias.html>
- [80] E. B. Kong and T. G. Dietterich, “Error-correcting output coding corrects bias and variance,” ” in *International Conference on Machine Learning*, pp. 313–321, 1995. [Online]. Available: <http://citeseer.ist.psu.edu/kong95errorcorrecting.html>
- [81] G. Valentini and T. G. Dietterich, “Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods,” *Journal of Machine Learning Research*, vol. 5, pp. 725–775, 2004.
- [82] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the margin: a new explanation for the effectiveness of voting methods,” ” in *Proceedings of 14th International Conference on Machine Learning*, pp. 322–330, 1997. [Online]. Available: <http://citeseer.ist.psu.edu/schapire97boosting.html>
- [83] T. K. Ho, “Data complexity analysis for classifier combination,” ” in *MCS ’01: Proceedings of the Second International Workshop on Multiple Classifier Systems*, pp. 53–67, 2001.
- [84] L. Kuncheva, C. Whitaker, C. Shipp, and R. Duin, “Is independence good for combining classifiers?” in *Proceedings of 15th International Conference on Pattern Recognition*, vol. 2, pp. 168–171, 2000.
- [85] L. I. Kuncheva, “Diversity in multiple classifier systems (editorial),” *Information Fusion*, vol. 6, no. 1, pp. 3–4, 2005.
- [86] L. I. Kuncheva and C. J. Whitaker, “Measures of diversity in classifier ensembles,” *Machine Learning*, vol. 51, no. 2, pp. 181–207, May 2003.
- [87] L. Kuncheva, M. Skurichina, and R. P. W. Duin, “An experimental study on diversity for bagging and boosting with linear classifiers.” *Information Fusion*, vol. 3, no. 4, pp. 245–258, 2002.

- [88] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *International Journal on Information Fusion*, vol. 6, no. 1, pp. 83–98, 2004.
- [89] A. Tsymbal, S. Puuronen, and D. W. Patterson, "Ensemble feature selection with the simple bayesian classification," *Information Fusion*, vol. 4, no. 2, pp. 87–100, 2003.
- [90] C. A. Shipp and L. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, no. 2, pp. 135–148, 2002.
- [91] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Information Fusion Journal (Special issue on Diversity in Multiple Classifier Systems)*, vol. 6, no. 1, pp. 5–20, 2005.
- [92] L. Kuncheva, "That elusive diversity in classifier ensembles," " in *First Iberian Conference on Pattern Recognition and Image Analysis, (IbPRIA 2003)*, pp. 1126–1138, 2003.
- [93] D. Tax and R. Duin, "Uniform object generation for optimizing one-class classifiers," *Journal of Machine Learning Research*, pp. 155–173, March 2002.
- [94] J. Platt, "Probabilities for SV Machines," in *Advances in Large Margin Classifiers*, A. J. Smola, P. J. Bartlett, B. Scholkopf, and D. Schuurmans, Eds., pp. 61–74, 2000.
- [95] S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," University of California, Irvine, Dept. of Information and Computer Sciences, 1998. [Online]. Available: <http://www.ics.uci.edu/~mlearn/MLRepository.html>
- [96] L. Cox, M. Johnson, and K. Kafadar, "Exposition of statistical graphics technology," " in *ASA Proceedings of the Statistical Computation Section*, pp. 55–56, 1982.

- [97] F. G. Yanowitz, "The Alan E. Lindsay ECG learning center in cyberspace," 2003. [Online]. Available: <http://medlib.med.utah.edu/kw/ecg/>
- [98] S. H. affiliated with the University of Florida, "The health illustrated encyclopedia," 2005. [Online]. Available: <http://www.shands.org/health/information/>
- [99] M. Kundu, M. Nasipuri, and D. K. Basu, "Knowledge-based ECG interpretation: a critical review," *Pattern Recognition*, vol. 33, no. 3, pp. 351–373, March 2000.
- [100] A. Taddei, M. Niccolai, M. Emdin, and C. Marchesi, "ABACUS: a knowledge-based system for the interpretation of arrhythmias in long term ECG," " in *Proceedings of Computers in Cardiology 1993*, pp. 887–890, Sept. 1993.
- [101] G. Bortolan, R. Degani, and W. Pedrycz, "A fuzzy pattern matching technique for diagnostic ECG classification," " in *Proceedings of Computers in Cardiology*, pp. 551–554, Sept. 1988.
- [102] W. Zong and D. Jiang, "Automated ECG rhythm analysis using fuzzy reasoning," " in *Computers in Cardiology 1998*, pp. 69–72, Sept. 1998.
- [103] M. Kundu, M. Nasipuri, and D. K. Basu, "A knowledge-based approach to ECG interpretation using fuzzy logic," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 28, no. 2, pp. 237–243, April 1998.
- [104] R. D. Throne, J. M. Jenkins, S. A. Winston, C. J. Finelli, and L. A. Dicarlo, "Discrimination of retrograde from antegrade atrial activation using intracardiac electrogram waveform analysis," *Pacing and Clinical Electrophysiology*, vol. 12, pp. 1622–1630, 1989.
- [105] S. C. Schuckers, X. Xu, M. E. Schuckers, and J. M. Jenkins, "Ventricular arrhythmia detection using time-domain template algorithms," " in *Proceedings of the IEEE 24th Annual Northeast Bioengineering Conference*, pp. 21–23, Apr. 1998.

- [106] J. Nadal and R. B. Panerai, "Classification of cardiac arrhythmias using principal component analysis of the ECG," " in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 13, pp. 580–581, Oct.-Nov. 1991.
- [107] R. Prado, I. Garcia, and P. Gomis, "Classification of high resolution ECG from chagasic patients with wavelet based bayesian models," " in *Computers in Cardiology 2001*, pp. 501–504, Sept. 2001.
- [108] P. de Chazal, M. O'Dwyer, and R. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196 – 1206, July 2004.
- [109] M. Risk, J. Sobh, and J. Saul, "Beat detection and classification of ECG using self organizing maps," " in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology society, 1997*, vol. 1, pp. 89–91, Oct.-Nov. 1997.
- [110] L. Vladutu, S. Papadimitiou, S. Mavroudi, and A. Bezerianos, "Ischemia detection using supervised learning for hierarchical neural networks based on kohonen-maps," " in *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 1688–1691, Oct. 2001.
- [111] M. Baig, A. Rasool, and M. Bhatti, "Classification of electrocardiogram using SOM, LVQ and beat detection methods in localization of cardiac arrhythmias," " in *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 1684–1687, Oct. 2001.
- [112] Y. H. Hu, S. Palreddy, and W. J. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach," *IEEE Trans. on Biomed. Eng.*, vol. 44, no. 9, pp. 891–900, Sep 1997.
- [113] Y. Sun, K. Chan, S. Krishnan, and D. Dutt, "Unsupervised classification of ECG beats using a MLVQ neural network," " in *Proceedings of the 22nd Annual*

- International Conference of the IEEE Engineering in Medicine and Biology Society, 2000*, vol. 2, pp. 1387–1390, July 2000.
- [114] R. Silipo and C. Marchesi, “Artificial neural networks for automatic ECG analysis,” *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1417–1425, May 1998.
 - [115] T. Stamkopoulos, K. Diamantaras, N. Maglaveras, and M. Strintzis, “ECG analysis using nonlinear PCA neural networks for ischemia detection,” *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 3058 – 3067, Nov. 1998.
 - [116] T. H. Yeap, F. Johnson, and M. Rachniowski, “ECG beat classification by a neural network,” ” in *Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1457–1458, Nov. 1990.
 - [117] Y. S. Tsai, B. N. Hung, and S. F. Tung, “An experiment on ECG classification using back-propagation neural network,” ” in *Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1463–1464, Nov. 1990.
 - [118] G. Oien, N. Bertelsen, T. Eftestol, and J. Husoy, “ECG rhythm classification using artificial neural networks,” ” in *IEEE Digital Signal Processing Workshop Proceedings, 1996*, pp. 514–517, Sept. 1996.
 - [119] C. Papaloukas, D. Fotiadis, A. Likas, and Michalis, “An ischemia detection method based on artificial neural networks,” *Artificial Intelligence in Medicine*, vol. 24, no. 2, pp. 167–178, Feb. 2002.
 - [120] C. Kramer, B. Mckay, and J. Belina, “Probabilistic neural network array architecture for ECG classification,” ” in *Proceedings of IEEE 17th Annual Conference on Engineering in Medicine and Biology*, vol. 1, pp. 807–808, Sept. 1995.
 - [121] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, and M. Strintzis, “ECG pattern recognition and classification using non-linear transformations and neural networks: A review,” *International Journal of Medical Informatics*, vol. 52, no. 1-3, pp. 191–208, Oct.-Dec. 1998.

- [122] P. Bozzola, G. Bortolan, C. Combi, F. Pinciroli, and C. BroHet, "A hybrid neuro-fuzzy system for ECG classification of myocardial infarction," " in *Computers in Cardiology 1996*, pp. 241–244, Sept. 1996.
- [123] S. Osowski and T. H. Linh, "ECG beat recognition using fuzzy hybrid neural network," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1265–1271, Nov. 2001.
- [124] M. Engin, "ECG beat classification using neuro-fuzzy network," *Pattern Recognition Letters*, vol. 25, no. 15, pp. 1715–1722, November 2004.
- [125] L.-Y. Shyu, Y.-H. Wu, and W. Hu, "Using wavelet transform and fuzzy neural network for VPC detection from the holter ECG," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1269 – 1273, July 2004.
- [126] T. Olmez, "Classification of ECG waveforms by using rce neural network and genetic algorithms," *Electronics Letters*, vol. 33, no. 18, pp. 1561–1562, Aug. 1997.
- [127] Z. Dokur and T. Olmez, "ECG beat classification by a novel hybrid neural network," *Computer Methods and Programs in Biomedicine*, vol. 66, no. 2-3, pp. 167–181, Sept. 2001.
- [128] T. Olmez and Z. Dokur, "Application of InP neural network to ECG beat classification," *Neural Computing & Applications*, vol. 11, no. 3-4, pp. 144–155, 2003.
- [129] J. Millet-Roig, R. Ventura-Galiano, F. J. Chorro-Gasco, and A. Cebrian, "Support vector machine for arrhythmia discrimination with wavelet-transform-based feature selection," *Computers in Cardiology*, pp. 407–410, 2000.
- [130] P. de Chazal and B. Celler, "Improving ECG diagnostic classification by combining multiple neural networks," " in *Computers in Cardiology 1997*, pp. 473–476, Sept. 1997.

- [131] I. Guler and E. D. Ubeyh, "ECG beat classifier designed by combined neural network model," *Pattern Recognition*, vol. 38, no. 2, pp. 199–208, February 2005.
- [132] R. Watrous and G. Towell, "A patient-adaptive neural network ECG patient monitoring algorithm," " in *Computers in Cardiology 1995*, pp. 229–232, Sept. 1995.
- [133] Y. H. Hu, S. Palreddy, and W. J. Tompkins, "Customized ECG beat classifier using mixture of experts," " in *Proceedings of the 1995 IEEE Workshop Neural Networks for Signal Processing*, pp. 459–464, Sept. 1995.
- [134] S. Palreddy, W. J. Tompkins, and Y. H. Hu, "Customization of ECG beat classifiers developed using SOM and LVQ," " in *IEEE 17th Annual Conference Engineering in Medicine and Biology Society*, vol. 1, pp. 813–814, Sept. 1995.
- [135] I. Christov, "Real time electrocardiogram QRS detection using combined adaptive threshold," *BioMedical Engineering OnLine*, vol. 3, no. 1, p. 28, 2004. [Online]. Available: <http://www.biomedical-engineering-online.com/content/3/1/28>
- [136] O. Fojt and J. Holcik, "Applying nonlinear dynamics to ECG signal processing," *IEEE Engineering in Medicine and Biology Magazine*, vol. 17, no. 2, pp. 96–101, March–April 1998.
- [137] M. Owis, A. Abou-Zied, A.-B. Youssef, and Y. Kadah, "Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 7, pp. 733–736, July 2002.
- [138] R. Gopalakrishnan, S. Acharya, and D. Mugler, "Automated diagnosis of ischemic heart disease using dilated discrete hermite functions," " in *Proceedings of the IEEE 30th Annual Northeast Bioengineering Conference*, pp. 97–98, April 2004.
- [139] Z. gen Zhang, H. zhong Jiang, D. fei Ge, and X. jian Xiang, "Pattern recognition of cardiac arrhythmias using scalar autoregressive modeling," *Fifth World*

- Congress on Intelligent Control and Automation (WCICA 2004)*, vol. 6, pp. 5545–5548, June 2004.
- [140] D. Ge and Z. Zhang, “Pattern recognition of cardiac arrhythmias based on multivariate autoregressive modeling,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 5, pp. V – 605–8, May 2004.
- [141] Z. Dokur, T. Olmez, and E. Yazgan, “Comparison of discrete wavelet and fourier transforms for ECG beat classification,” *Electronics Letters*, vol. 35, no. 18, pp. 1502–1504, Sept. 1999.
- [142] R. Mark and R. Wallen, “AAMI-recommended practice: Testing and reporting performance results of ventricular arrhythmia detection algorithms,” Association for the Advancement of Medical Instrumentation, Arrhythmia Monitoring Subcommittee, Tech. Rep. AAMI ECAR-1987, 1987.
- [143] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D. Tax, “PRTools4, A Matlab toolbox for pattern recognition,” Delft University of Technology, 2004. [Online]. Available: <http://www.prtools.org/>
- [144] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [145] D. M. J. Tax, “Data description toolbox (DD-Tools),” Delft University of Technology, 2004. [Online]. Available: http://www-ict.ewi.tudelft.nl/~davidt/dd_tools.html
- [146] Y. L. amd X. Yao, “Ensemble learning via negative correlation,” *Neural Networks*, vol. 12, pp. 1399–1404, 1999.
- [147] J. Rodriguez, A. Goni, and A. Illarramendi, “Real-time classification of ECGs on a PDA,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 1, pp. 23–34, March 2005.
- [148] “Colonoscopy,” Jackson Gastroenterology. [Online]. Available: <http://www.gicare.com/pated/epdgs19.htm>

- [149] “Fujinon EVE 200 endoscopes,” Fujinon Inc.(USA). [Online]. Available: <http://www.fujinon.com/product/medical/ed06.htm>
- [150] “Main products of fujinon-medical equipment,” Fujinon Inc.(USA). [Online]. Available: <http://www.fujinon.com/product/index.htm>
- [151] P. Li, K. L. Chan, S. M. Krishnan, and Y. Gao, “Detecting abnormal regions in colonoscopic images by patch-based classifier ensemble,” ” in *17th International Conference on Pattern Recognition (ICPR)*, pp. 774–777, August 2004.
- [152] P. Li, S. M. Krishnan, K. L. Chan, and Y. Gao, “Abnormal region detection in colonoscopic images using novelty detection technique,” ” in *IWAIT’2004, The International Workshop on Advanced Imaging Technology*, pp. 213–217, January 2004.
- [153] D. L. Pham, C. Xu, and J. L. Prince, “Current methods in medical image segmentation,” *Annual Review of Biomedical Engineering*, vol. 2, pp. 315–337, 2000.
- [154] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, August 2000.
- [155] S. X. Yu and J. Shi, “Segmentation given partial grouping constraints,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 173–83, Feb. 2004.
- [156] J. Carballido-Gamio, S. Belongie, and S. Majumdar, “Normalized cuts in 3-D for spinal MRI segmentation,” *IEEE Transactions on Medical Imaging*, vol. 23, no. 1, pp. 36–44, Jan. 2004.
- [157] X. Ren and J. Malik, “Learning a classification model for segmentation,” ” in *Ninth IEEE International Conference on Computer Vision (ICCV’03)*, vol. 1, pp. 10–17, October 2003.
- [158] K. Murphy, A. Torralba, and W. T. Freeman, “Using the forest to see the trees: a graphical model relating features, objects and scenes,” ” in *16th NIPS*, 2004.

- [159] N. Jojic, B. Frey, and A. Kannan, "Epitomic analysis of appearance and shape," " in *Proceedings of ICCV 2003*, pp. 34–41, October 2003.
- [160] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," " in *Proceedings. 2001 International Conference on Image Processing*, vol. 1, pp. 1046–1049, Oct. 2001.
- [161] N. V. Chawla and K. W. Bowyer, "Random subspaces and subsampling for 2-D face recognition," " in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 582–589, 2005.
- [162] S. Gupte, O. Masoud, R. F. K. Martin, and N. P. Papanikolopoulos, "Detection and classification of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 1, pp. 37–47, March 2002.
- [163] W. Jin, R. Wang, and L. Wu, "Patch-based natural object detection using CF*IRF," " in *ICME '04, 2004 IEEE International Conference on Multimedia and Expo*, vol. 3, pp. 1559–1562, June 2004.
- [164] T. Deselaers, D. Keysers, and H. Ney, "Discriminative training for object recognition using image patches," " in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 157–162, June 2005.
- [165] I. Ulusoy and C. M. Bishop, "Generative versus discriminative methods for object recognition," " in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pp. 258–265, 2005.
- [166] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "Segmenting, modeling, and matching video clips containing multiple moving objects," " in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 914–921, June-July 2004.
- [167] S. M. Krishnan and P. M. Y. Goh, "Quantitative parametrization of colonoscopic images by applying fuzzy technique," " in *Proceedings of the 19th Annual*

- International Conference of the IEEE Engineering in Medicine and Biology society*, vol. 3, pp. 1121–1123, Oct. - Nov. 1997.
- [168] S. M. Krishnan, X. Yang, K. L. Chan, S. Kumar, and P. M. Y. Goh, “Intestinal abnormality detection from endoscopic images,” ” in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 20, no. 2, pp. 895–898, 1998.
- [169] P. Wang, S. M. Krishnan, Y. Huang, and N. Srinivasan, “An adaptive segmentation technique for clinical endoscopic image processing,” ” in *Proceeding of the 24th International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 1084–1085, October 2002.
- [170] S. M. Krishnan, C. J. Yap, K. V. Asari, and P. M. Y. Goh, “Neural network based approaches for the classification of colonoscopic images,” ” in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology society*, vol. 3, pp. 1678–1680, Oct. - Nov. 1998.
- [171] M. M. Zheng, S. M. Krishnan, and R. Doraiswami, “Data fusion application for computer-assisted clinical endoscopic image analysis,” ” in *IEEE CCECE 2002. Canadian Conference on Electrical and Computer Engineering*, vol. 2, pp. 1116–1119, May 2002.
- [172] M. M. Zheng and S. M. Krishnan, “Decision support by fusion in endoscopic diagnosis,” ” in *The Seventh Australian and New Zealand Intelligent Information Systems Conference*, pp. 107–110, Nov. 2001.
- [173] M. M. Zheng, S. M. Krishnan, and M. P. Tjoa, “A fusion-based clinical decision support for disease diagnosis from endoscopic images,” *Computers in Biology and Medicine*, vol. 35, no. 3, pp. 259–274, March 2005.
- [174] M. P. Tjoa, S. M. Krishnan, C. J. Yap, S. Swaminathan, and P. Wang, “Application of artificial neural networks for classification of colonoscopic images,” ” in *2002 Asia-Pacific Conference on Circuits and Systems APCCAS '02*, vol. 2, pp. 227–230, Oct. 2002.

- [175] M. P. Tjoa and S. M. Krishnan, "Feature extraction for the analysis of colon status from the endoscopic images," *BioMedical Engineering Online*, vol. 2, no. 9, April 2003.
- [176] P. Wang, S. M. Krishnan, C. Kugean, and M. P. Tjoa, "Classification of endoscopic images based on texture and neural network," " in *Proceeding of the 23rd International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 3691–3695, October 2001.
- [177] S. A. Karkanis, D. K. Iakovidis, D. A. Karras, and D. E. Maroulis, "Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures," " in *Proceedings of 2001 International Conference on Image Processing*, vol. 2, pp. 833–836, Oct. 2001.
- [178] D. E. Maroulis, D. K. Iakovidis, S. A. Karkanis, and D. A. Karras, "Cold: a versatile detection system for colorectal lesions in endoscopy video-frames," *Computer Methods and Programs in Biomedicine*, vol. 70, no. 2, pp. 151–166, February 2003.
- [179] V. P. Magoulas, G. D. Plagianakos and M. N. Vrahatis, "Neural network-based colonoscopic diagnosis using on-line learning and differential evolution," *Applied Soft Computing*, vol. 4, no. 4, pp. 369–379, Feb. 2004.
- [180] G. D. Magoulas, V. P. Plagianakos, D. K. Tasoulis, and M. N. Vrahatis, "Tumor detection in colonoscopy using the unsupervised k-windows clustering algorithm and neural networks," " in *Proceedings of the Fourth European Symposium on Biomedical Engineering*, June 2004.
- [181] V. S. Kodogiannis, "Computer-aided diagnosis in clinical endoscopy using neuro-fuzzy systems," " in *Proceedings of 2004 IEEE International Conference on Fuzzy Systems*, vol. 3, pp. 1425–1429, July 2004.
- [182] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. D. Tzivras, "Computer aided tumor detection in endoscopic video using color wavelet features," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 3, pp. 141–152, September 2003.

-
- [183] S. Karkanis, G. D. Magoulas, D. K. Iakovidis, D. E. Maroulis, and N. Theofanous, "Tumor recognition in endoscopic video images using artificial neural network architectures," " in *26th EUROMICRO Conference, Medical Informatics*,, pp. 423–429, 2000.
- [184] S. A. Karkanis, D. K. Iakovidis, D. A. Karras, and D. E. Maroulis, "Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures," " in *IEEE International Conference in Image Processing (ICIP) Proceedings*, pp. 833–836, 2001.
- [185] W. K. Pratt, *Digital Image Processing*, 3rd ed., New York, 2001.
- [186] F. Roli, G. Fumera, and J. Kittler, "Fixed and trained combiners for fusion of imbalanced pattern classifiers," 2002. [Online]. Available: <http://citeseer.ist.psu.edu/roli02fixed.html>