

INCREMENTAL NONPARAMETRIC DISCRIMINANT
ANALYSIS BASED ACTIVE LEARNING AND ITS
APPLICATIONS

KSHITIJ DHOBLE

A THESIS SUBMITTED TO

AUCKLAND UNIVERSITY OF TECHNOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF COMPUTER AND INFORMATION SCIENCES (MCIS)

18TH MARCH 2010

KNOWLEDGE ENGINEERING AND DISCOVERY RESEARCH INSTITUTE (KEDRI)

PRIMARY SUPERVISOR: DR. PAUL S. PANG

SECONDARY SUPERVISOR: PROF. NIK KASABOV

ABSTRACT

Learning is one such innate general cognitive ability which has empowered the living animate entities and especially humans with intelligence. It is obtained by acquiring new knowledge and skills that enable them to adapt and survive. With the advancement of technology, a large amount of information gets amassed. Due to the sheer volume of increasing information, its analysis is humanly unfeasible and impractical. Therefore, for the analysis of massive data we need machines (such as computers) with the ability to learn and evolve in order to discover new knowledge from the analysed data.

The majority of the traditional machine learning algorithms function optimally on a parametric (static) data. However, the datasets acquired in real practices are often vast, inaccurate, inconsistent, non-parametric and highly volatile. Therefore, the learning algorithms' optimized performance can only be transitory, thus requiring a learning algorithm that can constantly evolve and adapt according to the data it processes. In light of a need for such machine learning algorithm, we look for the inspiration in humans' innate cognitive learning ability. Active learning is one such biologically inspired model, designed to mimic humans' dynamic, evolving, adaptive and intelligent cognitive learning ability.

Active learning is a class of learning algorithms that aim to create an accurate classifier by iteratively selecting essentially important unlabeled data points by the means of adaptive querying and training the classifier on those data points which are potentially useful for the targeted learning task (Tong & Koller, 2002). The traditional active learning techniques are implemented under supervised or semi-supervised learning settings (Pang et al., 2009). Our proposed model performs the active learning in an unsupervised setting by introducing a discriminative selective sampling criterion, which reduces the computational cost by substantially decreasing the number of irrelevant instances to be learned by the classifier.

The methods based on passive learning (which assumes the entire dataset for training is truly informative and is presented in advance) prove to be inadequate in a real

world application (Pang et al., 2009). To overcome this limitation, we have developed Active Mode Incremental Nonparametric Discriminant Analysis (aIncNDA) which undertakes adaptive discriminant selection of the instances for an incremental NDA learning. NDA is a discriminant analysis method that has been incorporated in our selective sampling technique in order to reduce the effects of the outliers (which are anomalous observations/data points in a dataset). It works with significant efficiency on the anomalous datasets, thereby minimizing the computational cost (Raducanu & Vitriá, 2008). NDA is one of the methods used in the proposed active learning model. This thesis presents the research on a discrimination-based active learning where NDA is extended for fast discrimination analysis and data sampling. In addition to NDA, a base classifier (such as Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN)) is applied to discover and merge the knowledge from the newly acquired data.

The performance of our proposed method is evaluated against benchmark University of California, Irvine (UCI) datasets, face image, and object image category datasets. The assessment that was carried out on the UCI datasets showed that Active Mode Incremental NDA (aIncNDA) performs at par and in many cases better than the incremental NDA with a lower number of instances. Additionally, aIncNDA also performs efficiently under the different levels of redundancy, but has an improved discrimination performance more often than a passive incremental NDA. In an application that undertakes the face image and object image recognition and retrieval task, it can be seen that the proposed multi-example active learning system dynamically and incrementally learns from the newly obtained images, thereby gradually reducing its retrieval (classification) error rate by the means of iterative refinement.

The results of the empirical investigation show that our proposed active learning model can be used for classification with increased efficiency. Furthermore, given the nature of network data which is large, streaming, and constantly changing, we believe that our method can find practical application in the field of Internet security.

ACKNOWLEDGMENT

This thesis would not have been possible without the support of many people. I wish to express my gratitude to my supervisors, Prof. Nik Kasabov and Dr. Paul S. Pang who have abundantly helped me and offered invaluable assistance, support and guidance.

Many thanks go in particular to Joyce D'Mello, for being always supportive.

Deepest gratitude are also due to all the members of KEDRI, without whose knowledge and assistance this study would not have been successful. I would also like to extend my thanks, to my stress-buster friends Lei Song and Gary Chen who have always been there for lunch, dinner, companionship at many outings and their dry humour about researcher's life.

I would also like to convey thanks to NICT, Japan for providing the scholarship and opportunity to work on their project. I would also like to acknowledge Auckland University of Technology for providing me a good study environment and facilities. Last, but not the least, I would like to express my love and gratitude to my parents for their understanding, support and endless love, through the course of my life.

PUBLICATIONS AND PRESENTATIONS

1) Pang, S., **Dhoble , K.**, Chen, Y., Kasabov, N., Ban, T., & Kadobayashi, Y. (2009, July). Active Mode Incremental Nonparametric Discriminant Analysis Learning. *In IMS '09: Proceedings of the Eighth International Conference on Information and Management Sciences*. (pp. 407-412). Kunming, China.

2) Pang, S., Chen, Y., Kasabov, N., & **Dhoble, K.** (2009). High Speed Algorithms for Outlier Detection and Classification over Huge-size Network Data Streams. Auckland, New Zealand: Auckland University of Technology, Knowledge Engineering and Discovery Research Institute (KEDRI), Research Report.

3) In preparation:

Dhoble , K., Pang, S., & Kasabov, N. (2009). Multi-example Image Retrieval on Active mode Incremental NDA Learning.

Contents

1	Introduction	1
1.1	Research Objective	4
1.2	Thesis Structure	5
2	Active Learning and Discriminant Analysis: A Review of Previous Works	7
2.1	Introduction	7
2.2	Approaches of Active Learning	8
2.3	Active Learning for Classification	9
2.3.1	Induction	13
2.3.2	Transduction	13
2.4	Subspace Analysis Review	15
2.4.1	Dimensionality reduction and Feature extraction	15
2.5	Active Learning: A review on Applications	18
2.5.1	Active Learning for Cyber Security	18
2.5.2	Active Learning in Bioinformatics	20
2.6	Summary	21

3	NDA Framework and System for Active Learning	22
3.1	Introduction	22
3.2	Nonparametric Discriminant Analysis (Batch NDA)	23
3.3	Incremental Nonparametric Discriminant Analysis (IncNDA)	24
3.4	Selective Sampling Criterion for Incremental Learning	25
3.4.1	Classification Accuracy Criterion (CAC)	25
3.4.2	Boundary Class Information Criterion (BCIC)	26
3.5	Active Learning Framework	29
3.5.1	Active mode IncNDA model (aIncNDA)	30
3.6	Summary	32
4	Discrimination Experiments on the Benchmark Datasets	33
4.1	Introduction	33
4.2	System Configuration	34
4.2.1	Software Configuration	34
4.2.2	Hardware Configuration	34
4.3	Experiment 1: CAC Criterion	37
4.3.1	Experimental Setup	37
4.3.2	Results	38
4.3.3	Discussion	39
4.4	Experiment 2: BCIC Criterion	39
4.4.1	Experimental Setup	40
4.4.2	Results	40
4.4.3	Discussion	44
4.5	Summary	46

5	Multi-example Image Retrieval Applications	47
5.1	Introduction	48
5.1.1	Related Researches and Motivation	49
5.2	Single example as an Image Query	51
5.3	Multiple example as an Image Query	51
5.4	Experiments and Discussion	53
5.4.1	Experimental Setup	53
5.4.2	Case Study 1: Face Image Retrieval	57
5.4.3	Case Study 2: Object Image Retrieval	61
5.4.4	Discussion	62
5.5	Summary	65
6	Conclusion and Future Works	66
6.1	Conclusion	66
6.1.1	Contributions	67
6.1.2	Limitations	67
6.2	Future Directions	68
6.2.1	Incremental evolving probabilistic spiking neural networks (pSNN) for active learning.	70
6.2.2	Quantum superposition as feature representation and feature selection for image classification and image associative memories.	70
	References	72

Appendices

A	Performance evaluation figures	81
----------	---------------------------------------	-----------

List of Figures

1.1	Active Learning Mechanism.	3
1.2	Passive Learning Mechanism.	4
2.1	Some of the commonly used Machine Learning Techniques for Classification tasks.	10
2.2	General Reinforcement Learning Schema: (1) Environment \rightarrow State (S_t) + Reward (r_t) \rightarrow Decision making process (2) Decision making process \rightarrow Action (a_t) \rightarrow Environment (3) Environment \rightarrow new State (S_{t+1}) + new Reward (r_{t+1})	12
2.3	General Schema of Inductive Inference based Classification Task: The inductive inference approach consists of two phases namely training and testing, where the classifier builds a general hypothesis based on training (labeled) data.	13
2.4	Transductive Inference approach. Adapted from Vapnik (2000). . . .	14
2.5	The Drug Discovery Cycle. Adapted from Warmuth et al. (2003). . .	20
3.1	Active mode IncNDA learning model (aIncNDA).	30
4.1	Performance evaluation of Classifiers.	35
4.2	The comparison of data distribution between the synthetic dataset and selected curiosity instances by proposed aIncNDA learning method. (a) The data distribution of the entire dataset; and (b) The data distribution of selected instance by aIncNDA.	41

4.3	The comparison of aIncNDA and IncNDA on the performance of incremental learning.	42
4.4	Comparison of aIncNDA and IncNDA on FR and FMA, (a) the performance of aIncNDA versus IncNDA on incremental learning; (b) the number of learned instances by aIncNDA at every learning stage.(using Discriminant Ratio based criterion)	43
5.1	Sample images from dataset 1: this dataset has face images with different facial expressions but have uniform illumination	55
5.2	Sample images from dataset 2: this dataset has face images with different facial expressions and lighting conditions	55
5.3	Sample AOLI wide baseline stereo object image. The combination of left-center and center-right images yields two pairs of 15 degree baseline stereo, and the left-right pair combination yields a 30 degree baseline stereo pair.(Geusebroek, Burghouts & Smeulders, 2005) . . .	56
5.4	Sample images from dataset 3: AOLI wide baseline stereo object images dataset	56
5.5	Dataset 2: Comparison on varying number of query images. (a) shows the retrieved images from two query images with one individual having glasses and the other having beard. In the retrieved images, we find images of same individuals along with other individuals having similar discriminative features such as beard or glasses. In (b), we add one more query with individual having both beard and sunglasses.	57
5.6	Images retrieved by single-example method and multi-example method from dataset 2. The images encapsulated in frames are incorrectly retrieved face images.	59
5.7	Performance evaluation (in terms of error rate versus number of iterations) of the proposed multi-image query based image retrieval on dataset 1. It can be seen that after 20 iterations the error rate stabilizes at 6.4%	60

5.8	Images retrieved by single-example method and multi-example method from dataset 3. The images encapsulated in frames are incorrectly retrieved object images.	61
5.9	Overall performance evaluation of the proposed multi-example method based image retrieval.	62
5.10	Dataset 2: Illumination Problem. (a) Image under uniform lighting/luminosity. (b) Image under lighting/luminosity focus from left side.	63
5.11	Dataset 2: Three-dimensional surface plot of images from Fig.5.10. The above figure shows how projection vectors are changed due to illumination/lighting conditions.	63
5.12	Example of image recognition using Affine Scale-Invariant Feature Transform (ASIFT) features. Amongst the key points found in both images using Harris-Affine method, 80 matches were found. The matching affine invariant key points between the two images are shown in the above figure.	64
A.1	D1 (Wisconsin) 2NN	82
A.2	D1 (Wisconsin) 5NN	82
A.3	D1 (Wisconsin) 7NN	82
A.4	D2 (Ionosphere) 2NN	83
A.5	D2 (Ionosphere) 5NN	83
A.6	D2 (Ionosphere) 7NN	83
A.7	D3 (Liver Disorder) 2NN	84
A.8	D3 (Liver Disorder) 5NN	84
A.9	D3 (Liver Disorder) 7NN	84
A.10	D6 (Iris) 2NN	85
A.11	D6 (Iris) 5NN	85

A.12 D6 (Iris) 7NN	85
A.13 D7 (Wine) 2NN	86
A.14 D7 (Wine) 5NN	86
A.15 D7 (Wine) 7NN	86
A.16 D8 (Heart) 2NN	87
A.17 D8 (Heart) 5NN	87
A.18 D8 (Heart) 7NN	87
A.19 D9 (Glass) 2NN	88
A.20 D9 (Glass) 5NN	88
A.21 D9 (Glass) 7NN	88
A.22 D12 (Face) 2NN	89
A.23 D12 (Face) 5NN	89
A.24 D12 (Face) 7NN	89

List of Tables

1.1	Comparisons between Active and Passive Learning Methodology. . . .	4
4.1	A comparison between SVM, LDA and IncNDA in terms of classification accuracy and confidence interval.	35
4.2	Incremental learning performed by NDA at different stages for Iris dataset. At every learning stage, 10% of the total data is provided. The percentage accuracy was calculated using leave-one-out cross validation.	36
4.3	Summary of evaluated UCI datasets.	38
4.4	Comparison between aIncNDA (Active) and IncNDA learning in terms of classification accuracy. Keys: DI = Dataset Index, SS = Samples selected by aIncNDA for learning, NN = Nearest Neighbour.	38
4.5	Comparison of aIncNDA versus IncNDA on Incremental Learning over 8 UCI benchmark datasets.	44
5.1	Performance evaluation using different number of query examples in multi-example method for (face image) dataset 1 and 2. Since the number of individuals in dataset 1 is 100, N/A denotes no data(individual image) available.	58
5.2	Overall percentage accuracy of single-example and multi-example method for face datasets.	60
5.3	Overall percentage accuracy of single-example method and multi-example method for dataset 3.	61

A.1 Overall percentage accuracy comparison between Euclidean distance (Euc.Distance) and Pearson's product-moment correlation coefficient (PMCC) (similarity metrics) method, using Single-example and Multi-example method for ALOI datasets.	90
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

List of Abbreviations

aIncNDA	Active mode Incremental Nonparametric Discriminant Learning
ALOI	Amsterdam Library of Object Images
ASIFT	Affine Scale-Invariant Feature Transform
BCIC	Boundary Class Information Criterion
CAC	Classification Accuracy Criterion
CBIR	Content based Image Retrieval
FDA	Fisher Discriminant Analysis
FMA	Face Membership Authentication
FR	Face Recognition
GMMs	Gaussian Mixture Models
HMMs	Hidden Markov Models
IDS	Intrusion Detection System
IncNDA	Incremental Nonparametric Discriminant Analysis
k-NN	k - Nearest Neighbor
LDA	Linear Discriminant Analysis
MDP	Markov Decision Processes
MeIR	Multi-example Image Retrieval
ML	Machine Learning
MPEG	Moving Picture Experts Group
NDA	Nonparametric Discriminant Analysis
PAC	probably-approximately-correct
PCA	Principal Component Analysis
PMCC	Pearson's Product-Moment Correlation Coefficient
SVM	Support Vector Machine
TED	Transductive experimental design
UCI	University of California, Irvine

Chapter 1

Introduction

From the beginning, a majority of the tools and machines designed by man have had their roots in the nature. Even during the stone-age period, weapons such as hunting blades were made in the form of animals' claw or teeth. The designs and functions shaped by nature are highly optimized and adaptive which has been and is being achieved after numerous generations of evolution. In this current era, which is ruled by Information Technology and Computer Science, we still look for an inspiration in the nature.

Learning is one such innate general cognitive ability which has empowered the living animate entities and especially humans with intelligence. It is obtained by acquiring new knowledge and skills which enable them to adapt and survive. With the advancement of technology, a large amount of information gets amassed. Due to the sheer volume of increasing information, its analysis is humanly unfeasible and impractical. Therefore, for the analysis of massive data we need machines (such as computers) with the ability to learn and evolve in order to discover new knowledge from the analysed data. In order to build such machines we look up to humans' innate cognitive learning ability for ideas. Active learning is one such biologically inspired model, designed to mimic humans' dynamic, incremental and intelligent cognitive learning ability.

Active learning plays a crucial role in classification (a stream of data mining); it requires fewer instances of labeled data for classification and thus solves the data scarcity problem. Active learning is a class of learning algorithms that aim to create

an accurate classifier by iteratively selecting essentially important unlabeled data points by the means of adaptive querying and training the classifier on those data points which are potentially useful for the targeted learning task. On the contrary, passive learning is static and non-adaptive, therefore it is completely dependent on the information that is already seen and hence it cannot adapt according to the new incoming data. Moreover, a passive learning model selects the data instances randomly resulting in reduced classifier accuracy and learning function (Pang et al., 2009). Random sampling is inefficient for classification in real world datasets since many such datasets have nonlinear distribution, resulting in the exclusion of a significant amount of informative instances (Tong & Koller, 2002). Due to active learning's adaptive and interactive nature it performs better than passive learning; especially where data is scarce and unlabeled. Furthermore, because of the selective sampling method incorporated in active learning, it only selects informative data instances based on a criterion. This results in the development of a better learning function and thus resulting in a more accurate classifier. The main objective of active learning model is to adaptively build an accurate classifier using the least number of data instances for training. Due to this nature of active learning, it is more efficient, precise and computationally less expensive than passive learning.

To further reduce the degree of the complexity in classification, several methods have been studied and implemented for feature selection. Discriminant Subspace Analysis is one such method that has been widely used for feature extraction and thus achieving dimensionality reduction. There are many such discrimination based methods such as Linear Discriminant Analysis (LDA) and Nonparametric Discriminant Analysis (NDA). In this thesis, we have used NDA method incorporated in the active learning model. These methods try to look for an optimal subspace in order to maximize the class separability. This class separability is achieved by simultaneously reducing the within class distance and increasing the between class distance. LDA achieves class separability through global eigenvectors whereas NDA achieves class separability through local eigenvectors (Pang et al., 2009). Datasets with higher dimensionality and fewer samples create singularity problem (Kuo & Landgrebe, 2004); however, NDA works on these types of datasets and retain important discriminant information after dimensionality reduction. In LDA, each class is assumed to be normally distributed with equal covariance matrix and these parametric distributions are optimally separated. Whereas, in NDA, class similarity is measured by the mean

distance of a certain number of neighboring samples and the separation of the classes is optimized based on this class similarity. Nonparametric methods such as NDA do not rely on an assumption that instances are drawn from a given probability distribution; therefore, they are more robust than parametric methods and work well on nonlinear datasets.

Active Learning vs. Passive Learning :

Active learning differs from passive learning where the former method attempts to select the most informative examples and train only those potentially useful for a particular learning task.

One analogy is that an active learner is a student who actively asks questions to a teacher, listens to the answers and asks further questions adaptively while a traditional passive learner is a student who listens to the teacher sitting in silence.

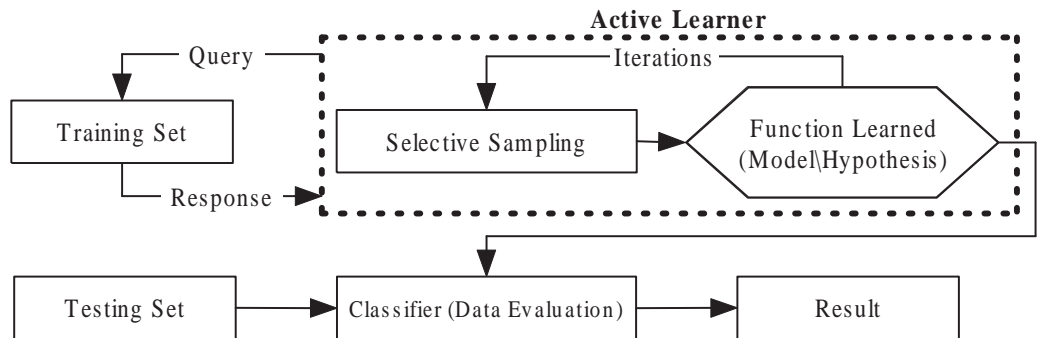


Figure 1.1: Active Learning Mechanism.

A passive learning system relies entirely on the previously gathered information, whereas an active learning algorithm has the capability of interacting with its environment in order to collect information and/or to select learning policy. Active learning systems produce improved generalization, reduce data costs and are most useful where data is expensive and computation is cheap (Symons et al., 2006). There are three major recognized approaches to the implementation of active learning: goal-driven learning, reinforcement learning and querying.

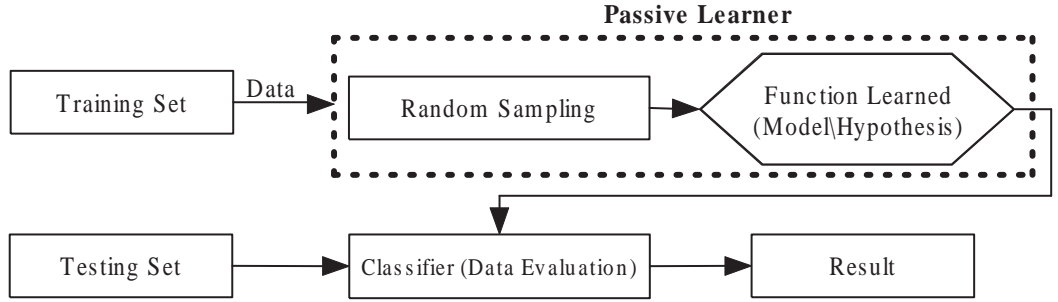


Figure 1.2: *Passive Learning Mechanism.*

The active learning system comprises two parts: a learning engine and a selection engine (Nguyen & Smeulders, 2004). During the iterations, the learning engine uses a supervised learning approach to train the classifier on labeled examples. The selection engine then selects a sample from the unlabeled dataset and requests a human expert to label the sample before passing it to the learning engine. The major goal is to achieve the best possible classifier within a reasonable number of iterations.

Table 1.1: *Comparisons between Active and Passive Learning Methodology.*

No	Active Learning	Passive Learning
1	Requires small training set	Requires large training set
2	Selective sampling of examples	Random sampling of examples
3	Few labeled examples required	Labeled examples only
4	Low Concept Drift Risk	High Concept Drift Risk
5	Reduces manual labeling cost	Has manual labeling cost

1.1 Research Objective

The recent research and investigation in active learning methods suggest many application areas such as cyber security (Almgren & Jonsson, 2004), bioinformatics (Liu, 2004; Warmuth et al., 2003), image processing (Wu, Tian & Huang, 2000;

Weber, Welling & Perona, 2000) and other domains involving pattern recognition tasks. Active learning is a class of machine learning algorithm which performs brain-like computational analysis and modeling. Its learning mode is similar to that of humans, where it dynamically and adaptively learns from previous and incoming data/information and accordingly builds an evolving model for tasks such as classification and predictions. Active learning models inspiration comes from the cognitive learning ability of the human brain, whereas spiking neural network models inspiration comes from the spiking processes in biological neurons. Both the algorithms have their own pros and cons. However, since active learning has been recently explored, there is still a lot to be further investigated.

This research aims to employ machine learning and data mining techniques to automate, speed up and increase the reliability of classification tasks. As part of the research, we have introduced criteria in the discrimination based active learning model with an incremental learning that integrates new informative model obtained from the incoming data and compared against other discriminant methods.

1.2 Thesis Structure

The thesis is structured as follows:

Chapter 2 contains a brief review of previous research on active learning approaches. In the review, different selective sampling methods used in active learning framework are considered. Also, various discriminant analysis techniques are underlined. Furthermore, the application of active learning in various domains is highlighted.

Chapter 3 discusses the proposed novel framework and system used for implementing active learning. The chapter begins with an introduction to Nonparametric Discriminant Analysis (NDA) and Incremental Nonparametric Discriminant Analysis (IncNDA), which are two of the components of selective sampling engine used in active learning. Also, two proposed selective sampling criteria which are a part of selective sampling engine are discussed at length.

Chapter 4 confers the experiments and analysis on the benchmark datasets used for investigating the efficacy of the Active mode Incremental Nonparametric Discriminant Analysis (aIncNDA), along with system specifications used for implementation and the two selective sampling criteria that were proposed for aIncNDA based active learning.

Chapter 5 presents the application of the novel aIncNDA based active learning for image recognition and retrieval tasks. This chapter commences with a brief introduction on previously used image recognition and retrieval techniques and the problem specifications encountered followed by the active learning (aIncNDA) framework and method used for image recognition and retrieval tasks. The chapter concludes with the experimental results and discussion about the advantages and limitations unearthed in the proposed aIncNDA active learning method.

Chapter 6 concludes the presented thesis along with suggestions for future work directions.

Chapter 2

Active Learning and Discriminant Analysis: A Review of Previous Works

2.1 Introduction

In the previous chapter, we introduced active learning and the difference between active and passive learning. This chapter introduces the different approaches and flavors of active learning along with their applications in different domains. It also presents a brief review of previous studies on active learning approaches. In the review, the different selective sampling methods used in active learning framework are highlighted. Also, various discriminant analysis techniques have been discussed. Furthermore, the applications of active learning in various domains such as internet security and bioinformatics have also been discussed.

The concept of active learning has only been explored recently. Previously, NDA has not been used as a selective sampling engine in active learning model. Most of the previous works were carried out on a semi-supervised learning; whereas, our model aim for unsupervised one-pass active learning. Some of the commonly used active learning directions/approaches that have been widely implemented are: Pool-based, Membership queries, Uncertainty Sampling, Information-based loss functions and Query by committee.

2.2 Approaches of Active Learning

There are varieties of selective sampling approaches used in active learning models. One of the approaches is *Pool-based active learning* which is the most commonly used query refinement scheme. However, it suffers from a multiple drawbacks; since most of the pool-based active learning iteratively selects samples from the pool which is informative or irrelevant, it becomes computationally intensive if the pool is small. Under such conditions this technique is ineffective in reducing the error rate (Ling & Du, 2008). Moreover, selecting the samples to be included in the pool itself is a time consuming process.

Another selective sampling approach is *Membership query* which selects samples directly from the dataset for the purpose of querying for labels. Membership query scheme does not have the drawbacks posed by the pool-based scheme. It also reduces the predictive error rapidly and is less computationally intensive compared to the pool-based active learning (Ling & Du, 2008).

Feature-value acquisition at cost is another approach where the active classifier has to obtain the values of the unlabeled data at some cost which is calculated using probably-approximately-correct (PAC) model. The calculation is based on the cost required to obtain additional values versus the penalty imposed on an inaccurate classification. It is one of the direction which has been used by Greiner, Grove and Roth (2002); Kapoor and Greiner (2005), where the cost of the feature value acquisition cannot exceed the budget which has been previously decided.

Transductive experimental design (TED) is another novel approach which was proposed by Yu, Bi and Tresp (2006) and is used to directly reduce the assessed uncertainty of the predictions on given unlabeled data, and thus effectively explored the information of unlabeled data in active learning (Zhang, He, Rey & Jones, 2007). Greedy algorithm performs problem solving (generally by combining user-given heuristic procedures) by making a local optimal choice for every example/instance in order to find the global optimum solution (Cormen, Leiserson & Rivest, 1990). But this technique faces the NP-hard problem as in (Yu et al., 2006; Zhang et al., 2007) and as a solution Kai, Zhu, Xu and Gong (2008) has proposed the use of non-greedy approach.

Clustering (Nguyen & Smeulders, 2004) and *Batch mode active learning* have been

used by Hoi, Jin and Lyu (2006) and is commonly employed as the Pool-based active learning approach. These are some of the other flavors of active learning which aims at decreasing the redundancy amongst the selected examples therefore providing more unique examples for the refinement of classifiers.

Lastly, *Query by Committee* technique used by Melville and Mooney (2004) is an effective approach where selective sampling is based on the disagreement amongst an ensemble of hypotheses for selecting data for labeling. Some of the commonly used ensemble with active learning includes techniques such as Bagging and Boosting (Abe, Zadrozny & Langford, 2006).

Our proposed model closely matches the membership query approach, since NDA (which has been used as a selective sampling engine in the active learning model) selects samples directly from the dataset. Incorporation of active learning with support vector machine has been commonly used especially in the field of bio-informatics (Danziger, Zeng, Wang, Brachmann & Lathrop, 2007; Liu, 2004; Warmuth et al., 2003), cyber security (Long, Yin, Zhu & Zhao, 2008; Almgren & Jonsson, 2004), multimedia information retrieval (Hoi & Lyu, 2005; Kherfi, Ziou & Bernardi, 2004; Huang et al., 2008) and text categorization (Hoi et al., 2006; Tong & Koller, 2002). However, a majority of these have made use of the pool-based technique which suffers from many drawbacks stated previously; therefore, it is recommended that although an incorporation of active learning with SVM is good, other approaches such as membership querying or batch mode active learning should be used as they negate the drawbacks introduced by pool based learning. It has been observed by Dasgupta and Hsu (2008) that both margin-based and cluster-adaptive sampling outperformed random sampling.

2.3 Active Learning for Classification

Several investigations of active learning have been done for classification tasks in a supervised or semi-supervised setting. Also, there are many methods which have been developed using active learning for the purpose of classification, regression (Sugiyama & Nakajima, 2009; Schein, 2005) and function optimization (Tong, 2001). There are many active learning - based classification algorithms such as the Query

by Committee algorithm (Melville & Mooney, 2004; Seung, Oppor & Sompolinsky, 1992; Freund, Seung, Shamir & Tishby, 1997), that (as explained in the previous section) queries the samples having higher disagreement amongst the given set or committee of classifiers. Tong (2001), states that the general form of this algorithm along with classifiers have been applied in several domains for the task of sampling from a probabilistic distribution. Tong (2001) also mentions that classifiers such as Naive Bayes have also been used but their performance is not as optimal as those of discrimination based classifiers such as Support Vector Machines (SVM), especially in the text classification domain (Siolas & Buc, 2000; Tong & Koller, 2002).

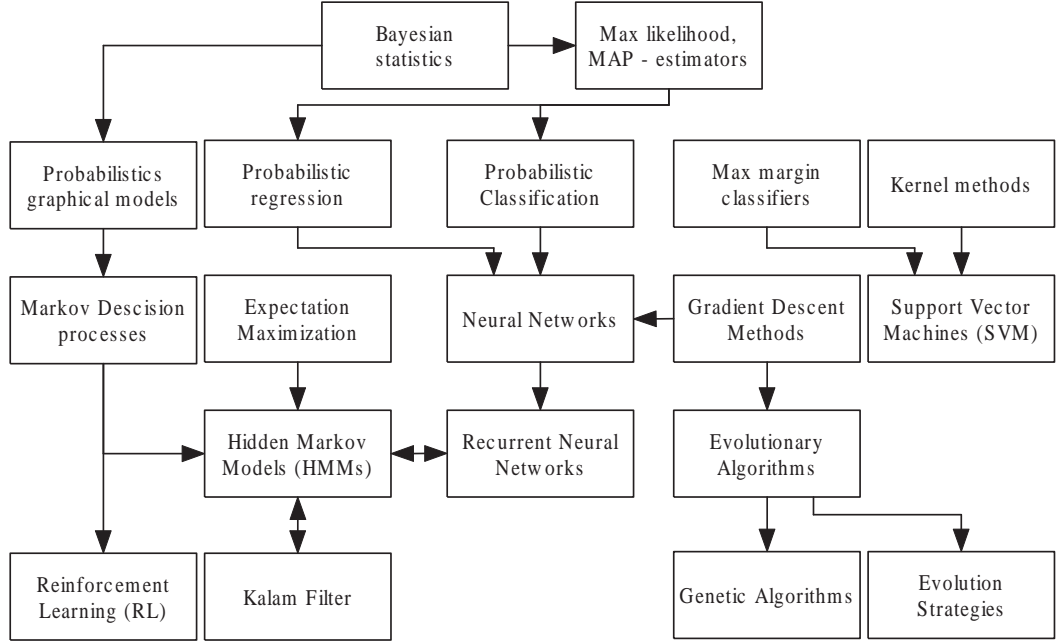


Figure 2.1: Some of the commonly used Machine Learning Techniques for Classification tasks.

In our model, we have used incremental nonparametric discriminant analysis (IncNDA) along with support vector machine. In aIncNDA, both IncNDA (used as a component of selective sampling engine) and SVM (used as a base classifier) being discriminant techniques complement each other and perform optimally through co-operative learning as compared to using k-NN classifier. In chapter 4, it can be seen that the IncNDA-SVM combination performs better than IncNDA-k-NN on the benchmark datasets.

Another strand of active learning is uncertainty sampling introduced by Lewis and Gale (1994), where instances which the classifier is most uncertain about are selected for learning. Tong (2001), states that Lewis and Gale (1994) successfully applied their uncertainty sampling technique in text domain using logistic regression. Active learning is very much similar to uncertainty sampling, where the learner uses a criterion for determining the relevance or informativeness of a particular instance or a chunk of instances.

To the best of our knowledge, not much study has been done on unsupervised active learning. Most of the work done on active learning involve supervised or semi-supervised learning (by using the *relevance feedback* from the user). However, in our work we aim for an unsupervised active learning model by introducing selection criterions that replaces the oracle/person. According to Tong (2001), such unsupervised techniques will prove to be beneficial in robotics, especially in the navigational system that is based on active learning system.

Reinforcement learning (Kaelbling, Littman & Moore, 1996; Tong, 2001) is one of the major areas of machine learning that cannot be clearly classified as supervised or unsupervised learning. It is used to solve a class of problems commonly known as Markov Decision Processes (MDP). Unlike supervised learning, the correct input and output are never provided. There is a classical exploration/exploitation trade-off in reinforcement learning where a reward is presented on positive outcome. In case of exploration, the reinforcement learning tries to find in space an appropriate behavior by refining the parameters of the whole model in order to improve the reward; whereas, in exploitation setting, as the reinforcement the learner tries to improve; its reward is based on the existing model (Kaelbling et al., 1996).

Active learning is similar to the exploration mode of reinforcement learning where the learner tries to find a new way to improve performance by adaptively changing its current model by learning as much as possible about the domain. Therefore, it can be said that in a way, active learning takes into consideration both spatial and temporal features of the current and incoming data in order to obtain an updated and optimized model.

Before we go ahead to the section that revises active learning and its application, a brief introduction about classification and its types will be provided. In our daily life, humans perform classification tasks effortlessly and naturally: recognizing our

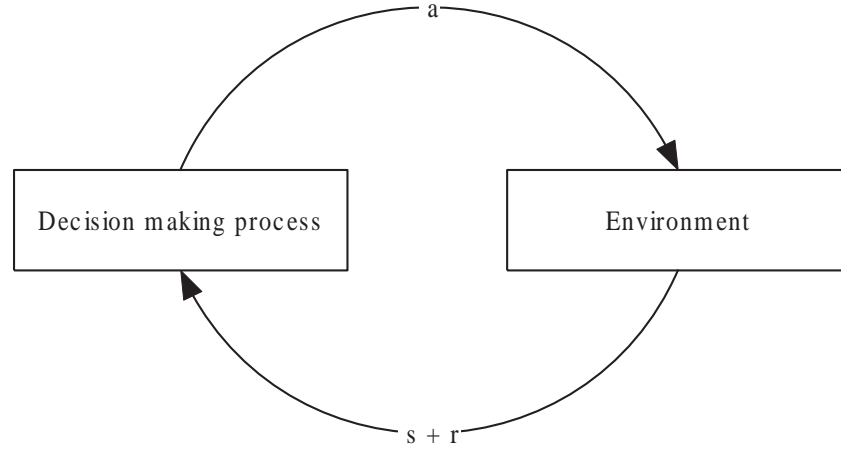


Figure 2.2: General Reinforcement Learning Schema: (1) Environment \rightarrow State (S_t) + Reward (r_t) \rightarrow Decision making process (2) Decision making process \rightarrow Action (a_t) \rightarrow Environment (3) Environment \rightarrow new State (S_{t+1}) + new Reward (r_{t+1})

friends' or relatives' face or voice in a crowd, detecting a particular ingredient in our food just by the sense of smell or taste, or recognizing a particular constellation or celestial object like comet in the sky. Classification is also frequently used in scientific endeavors for a varied range of tasks such as speech and image recognition (Kasabov, 1996), gene expression data analysis (Kasabov, Middlemiss & Lane, 2003), intrusion detection (Long et al., 2008) etc. The goal of classification is to make a classifier automatically classify/categorize given (new/incoming) instances, by making the classifier learn from historical data.

In general, classification techniques can be categorized into Induction and Transduction types. Until now the majority of methods in the machine learning (ML) domain that have been implemented either use the inductive or transductive approach for classification task. The inductive classification tasks, as the name suggests, formulate tentative hypotheses (learning function) from the training data in order to predict/classify the testing data (Pang & Kasabov, 2004) resulting in a more general solution; whereas, in transductive classification tasks, both the training and testing data are utilized in order to obtain a local solution i.e. for a particular (new) instance. Pang and Kasabov (2004) state that transductive - based methods prove to be more appropriate for medical or clinical (data) application. In the following discussion, a more detailed explanation on these two classification tasks is presented.

2.3.1 Induction

In statistical inference, induction or inductive inference based classification is most standard and commonly used classification method. The inductive inference based classification generally consists of two phases. The first phase consists of training where the classifier is provided (historical data) instances to learn (train) from, having identically distributed data $\{d_1 \dots d_n\}$ residing in space U . These data instances also have labels $\{l_1 \dots l_n\}$ where the possible set of labels L is discrete. These labeled instances are generally called training data or historical data. This training data is given to the classifier to build a (model) learning function $f : U \rightarrow L$ which completes the training phase (Joachims, 1999).

In the testing phase, the trained classifier f is then used to automatically classify new (unlabeled) instances $\{d'_1 \dots d'_n\}$. The new unlabeled data provided for testing is identically distributed and has the same probabilistic distribution (as in training phase). The performance of the classifier is measured on how accurately it is able to classify the (new/unlabeled) instances provided in the testing phase (Pang & Kasabov, 2004).

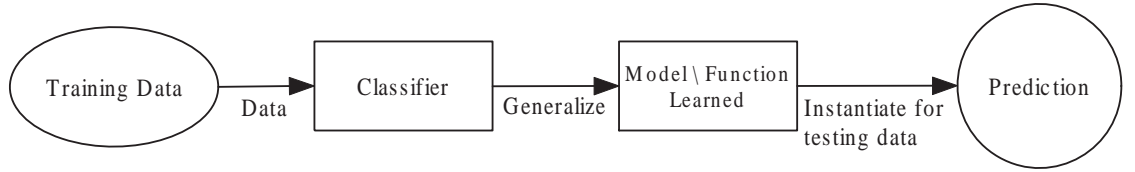


Figure 2.3: General Schema of Inductive Inference based Classification Task: The inductive inference approach consists of two phases namely training and testing, where the classifier builds a general hypothesis based on training (labeled) data.

2.3.2 Transduction

Transduction or transductive inference based reasoning is another approach for classification tasks. Compared to inductive inference approach where reasoning (function) is obtained from training data, transductive inference approaches reasoning is observed from specific training and testing data. In transductive inference setting the testing data $\{d'_1 \dots d'_n\}$ is known but still unlabeled. Also, we are provided with

identically distributed (training) data $\{d_1 \dots d_n\}$ residing in space U . The goal of the classifier is to simply provide labels $\{l'_1 \dots l'_n\}$ to the unlabeled instances.

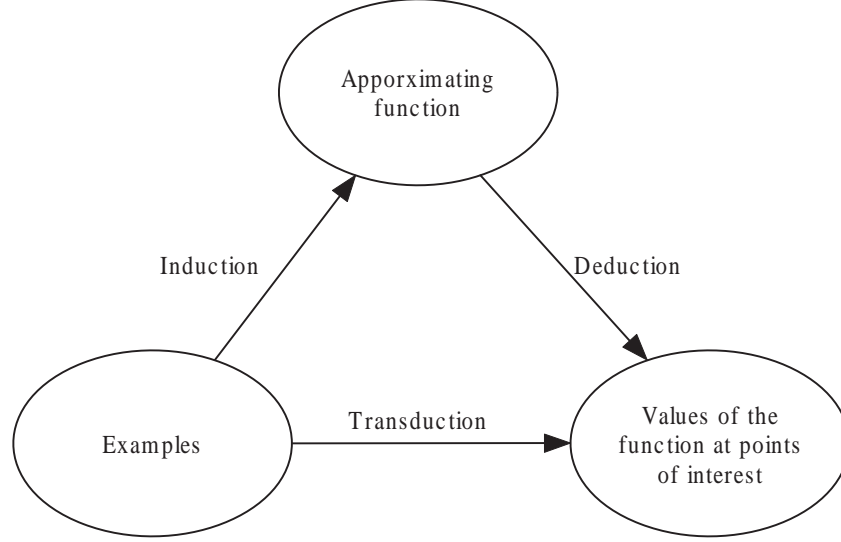


Figure 2.4: *Transductive Inference approach. Adapted from Vapnik (2000).*

It can be noted that even if the transductive tasks are solved using inductive method, there are many transductive based algorithms (Joachims, 1999; Vapnik, 1998; Pang & Kasabov, 2004) which take an advantage of the available unlabeled testing data and show improved performance over the standard inductive inference - based methods. In the article by Pang and Kasabov (2004), the authors have carried out the performance evaluation of inductive and transductive inference based methods on a medical dataset which shows that transductive inference methods (which takes the local information into consideration) performs better than inductive inference approach. However, the authors state that transductive approach is efficient only on small sized datasets. Therefore, we are hoping to address this problem in an active learning way.

In many of the supervised or semi-supervised tasks, labeling data for training set can prove to be time consuming and costly process. Therefore, finding methods to reduce the time and cost by selecting fewer labels will prove to be beneficial. Active learning is one of the methods that can be used to address this problem. Most of the supervised methods need labeled data for the training set; however, an active learning technique can help in reducing the number of labeled instances required for

training by choosing only those instances that will be informative for training the classifier by the means of selective sampling.

In our novel unsupervised active learning framework, we have incorporated Incremental Nonparametric Discriminant Analysis (IncNDA) as a part of selective sampling engine for the purpose of feature selection and incremental learning. In the following section, a brief review on various subspace analysis methods is presented and the decision regarding the selection of nonparametric discriminant analysis (NDA) method over other discriminative techniques is justified.

2.4 Subspace Analysis Review

Discriminative and informative learning are two different approaches used for classification in pattern recognition (Rubinstein & Hastie, 1997). The informative learning strategy (as in Gaussian mixture models (GMMs) and Hidden Markov Models (HMMs)) utilizes information about the classes by concurrently taking all the classes information (such as class density) into consideration; whereas, the discriminative learning strategy (as in SVM) makes use of discriminative information between the various classes or instances by simultaneously considering all the classes (Wang & Paliwal, 2002). We will be focusing on the discriminative analysis techniques in the following section.

2.4.1 Dimensionality reduction and Feature extraction

Feature extraction and dimensionality reduction is a common pre-processing step in classification tasks; where both parametric and nonparametric techniques have been implemented. The subspace analysis includes techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). A detailed review of these methods is presented below.

Principal Component Analysis (PCA)

Pearson (1901) introduced the *Principal Component Analysis* (PCA) and since then it has been widely used for data analysis and dimension reduction. PCA optimally

transforms the given data (without information loss) in order to achieve the largest covariance variation such that each of the largest variation lies on an axis that is known as principal axis. PCA works under the assumption that along these principal axes where the variation is the largest it contains most information about the classes. In the transformed PCA space, there is no redundant information since all the principal components are orthogonal to each other (due to linear combination of original values) (Pearson, 1901).

PCA takes into consideration global information about the dataset where the principal axis can be calculated using the global covariance matrix formula presented below (Wang & Paliwal, 2002):

$$\hat{S} = \frac{1}{N} \sum_{j=1}^{C_N} \sum_{i=1}^{C_j} (x_{ji} - \bar{\mu}) (x_{ji} - \bar{\mu})^T, \quad (2.1)$$

where C_N is number of classes, C_j represents the number of instances in class j , $\bar{\mu}$ is the global mean of all instances, $N = \sum_{j=1}^{C_N} C_j$ represents the i^{th} observation from class j .

Therefore according to Wang and Paliwal (2002),

$$\hat{S}T_1 = \hat{\lambda}_i T_i \{i \in 1, \dots, m\}, \quad (2.2)$$

where, $\hat{\lambda}_i$ is the i^{th} largest eigenvalue of \hat{S} , m represents the leading eigenvectors of \hat{S} for principal axis T_1, T_2, \dots, T_m . For a more detailed information on PCA refer to Pearson (1901).

Although PCA retains the subspace having the greatest covariance, it is not suitable for classification task since it also retains noise and works under the assumption that the underlying probability distribution for the given datasets is linear.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis(LDA), also known as *Fisher Discriminant Analysis* (FDA), seeks optimal projection of training data by simultaneously calculating the within and between - class covariance matrix. Pang, Ozawa and Kasabov (2005)

states that LDA works by seeking efficient discrimination; while PCA works by seeking efficient representation.

The within-class covariance matrix S_W and between-class covariance matrix S_B according to Wang and Paliwal (2002) can be represented as:

$$S_W = \sum_{j=1}^{C_N} \sum_{i=1}^{C_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T, \quad (2.3)$$

and

$$S_B = \sum_{j=1}^{C_N} C_j (\mu_j - \mu)(\mu_j - \mu)^T, \quad (2.4)$$

where, μ is the global mean, μ_j is mean of class j , C_N is the number of classes, C_j represents all the samples belonging to class j . The (LDA) discriminant eigenspace model Ω_{LDA} can be obtained through eigenvalue decomposition of the within and between - class covariance matrix which can be represented as:

$$\Omega_{LDA} = tr(S_W^{-1}S_B). \quad (2.5)$$

Both the above PCA and LDA techniques consider global information and are parametric in nature. However, most of the real world dataset are nonparametric in nature; therefore, techniques such as nonparametric discriminant analysis (NDA) are better suited for real world dataset (Kuo & Landgrebe, 2004; Raducanu & Vitriá, 2008).

Nonparametric Discriminant Analysis works by classifying objects into mutually exclusive and exhaustive groups based on a set of measurable object's features. Parametric Discriminant Analysis has an innate problem which originates from the parametric characteristics of the scatter matrix, in which the instances distribution for all the class are assumed to be normal distribution. So it tends to suffer in the case of non-normal distribution (Raducanu & Vitriá, 2008).

According to Raducanu and Vitriá (2008), nonparametric methods are distribution

free which considers fewer assumptions than the parametric methods, allowing wider ranges of applications, especially in those cases where less is known about the domain of application. Moreover, since non-parametric methods such as NDA rely on fewer assumptions, they are more robust in nature.

In the next chapter we have presented the detailed workings of a nonparametric method, which in our case is NDA and have also shown as to how we have incorporated the incremental version (Raducanu & Vitriá, 2008) into our novel active learning framework.

2.5 Active Learning: A review on Applications

In this section, we will be reviewing the application of active learning in various domains such as bioinformatics and cyber security.

2.5.1 Active Learning for Cyber Security

The application of *active learning* in the field of internet security is not mature and a lot remains to be explored. Currently, there have been very few applications in fields such as active response intrusion detection and malicious computer software detection. An intrusion detection system (IDS) detects various types of attacks. The two general methods used for intrusion detection are signature - based and anomaly - based. In signature - based method the IDS looks for a signature match for detecting an intrusion/attack. The anomaly - based method looks for new types of intrusion based on the *abnormal behavior*. However, the anomaly - based method experiences issues such as false alarms due to the fact that it considers any activity other than the regular as an intrusion.

Active learning can play a major role in efficient signature match and lowering of false alarms in anomaly based intrusion detection. Though much research has been done on application of machine learning in intrusion detection however active learning has rarely been employed. Furthermore there are many issues related to active learning

which needs to be resolved. Conversely, the problems related to active learning can be minimized using alternative sampling methods of the unlabeled data points.

Intrusion Detection System (IDS) plays vital role of detecting various kinds of attacks. The main purpose of IDS is to find out intrusions among normal audit data and this can be considered as classification problem. The two basic methods of detection are signature based and anomaly based (Ling & Du, 2008). The signature-based method, also known as misuse detection, looks for a specific signature to match, signaling an intrusion (Almgren & Jonsson, 2004). They can detect many or all known attack patterns, but they are of little use for as yet unknown attack methods. Most popular intrusion detection systems fall into this category.

Another method to intrusion detection is called anomaly detection. Anomaly detection applied to intrusion detection and computer security has been an active area of research since it was originally proposed by Denning (1987). The advantage of anomaly detection is that it can recognize new types of intrusions based on those patterns or behavior that are different from normal usage. The anomaly detection based algorithm is given normal data for training and new data for testing, where the algorithm aims to find if the given test data has anomalous or normal behavior/pattern. However, previously unseen but authentic system behaviors or pattern on which the algorithm has not been trained are recognized as anomalies and therefore labeled as possible intrusions causing the anomaly detection algorithm to suffer from high false alarm rate.

Long et al. (2008) has used a novel approach which uses active cost-sensitive learning method for intrusion detection. The author utilizes query by committee sampling method with the aim to reduce the burden of labeling data for constructing the intrusion detection classifier with the least misclassification cost. Introduction of active learning for intrusion detection using history data showed improved performance on the KDDCUP 99 dataset. In the traditional passive approach as stated by Lee, Stolfo and Mok (1999), high quality history data requires heavy labor of experts or expensive monitoring process. However, this problem as seen in the article by Long et al. (2008) is resolved by introducing active learning method.

2.5.2 Active Learning in Bioinformatics

Introduction of active learning in the bioinformatics domain has helped to accelerate several major studies such as biomolecular structure prediction, gene finding (Danziger et al., 2007), genomics and proteomics, cancer classification (Liu, 2004) and drug discovery (Warmuth et al., 2003).

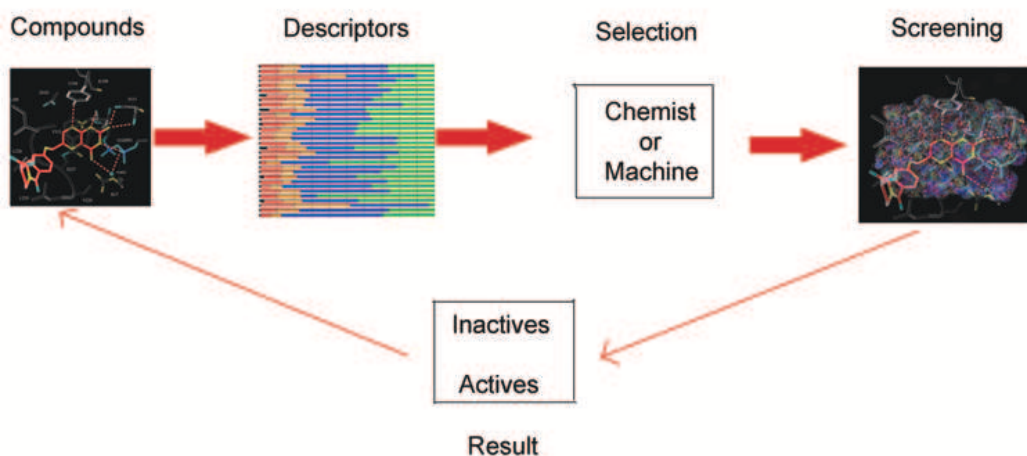


Figure 2.5: *The Drug Discovery Cycle. Adapted from Warmuth et al. (2003).*

In Fig. 2.5 it can be observed how Warmuth et al. (2003) used active learning technique in the selection phase to find out if the given drug compound is active or inactive against a biological target. Machine learning techniques are used in bioinformatics since they are suitable for understanding and discovering knowledge from the complex relationship present in the clinical or medical data (Pang & Kasabov, 2004). Active learning technique has been successfully used for solving problems in bioinformatics for tasks such as facilitating drug development, understanding tumour behaviour etc. Utilizing machine learning in bioinformatics is an efficient and inexpensive way of solving problems because the actual process such as in drug discovery, involves many iterations of biochemical testing for identifying the compounds active against the targeted molecule or site. Machine learning reduces these iterations by predicting which possible compounds are potentially active (or relevant) and based on these findings the biochemist/biologist can carry out biochemical tests on those potential compounds predicted by machine learning algorithms.

Besides bioinformatics and internet security, active learning based techniques also

have usage in image/multimedia recognition and retrieval. The application of active learning for image recognition and retrieval will be discussed in Chapter 5.

2.6 Summary

In this chapter, previous works on active learning have been discussed along with the different active learning approaches, and its applications in various domains. We have also reviewed various discriminative and classification techniques which have been used over the years. In the following chapter we will present a detailed study of the techniques and criteria incorporated in our method, followed by the workings of our novel active learning framework.

Chapter 3

NDA Framework and System for Active Learning

3.1 Introduction

This chapter presents experimental research on discrimination-based active learning for classification tasks. In this experimental research, a quantitative research methodology has been used for the creation, the analysis and the performance evaluation of the proposed model/framework. Intrusion detection system (IDE) relates to many issues, one of them being detecting intrusion in real time by analyzing large volume of the streaming data. The proposed novel active learning method for intrusion detection uses classifier such as SVM or k-NN in combination with incremental nonparametric discriminant analysis for classification. SVM in previous research works have proven to be more accurate and efficient in numerous real-world learning (Tong & Chang, 2001) and nonparametric discriminant analysis (NDA and IncNDA) for selective sampling since it is more robust than parametric discriminant analysis (Kuo & Landgrebe, 2004). Use of incremental learning is essential since it's not computationally feasible to compute from scatter matrix again. Therefore incremental learning is beneficial since it constantly updates the eigenspace (Raducanu & Vitriá, 2008) reducing the computational cost and time. In the following, a detailed working of Nonparametric Discriminant Analysis (Batch NDA) and Incremental Nonparametric Discriminant Analysis (IncNDA) has been presented.

3.2 Nonparametric Discriminant Analysis (Batch NDA)

Assuming that the data samples we have belong to N classes, therefore C_i represents samples belonging to one of the class i , $i = 1, 2, 3, \dots, N$. Therefore, the within class covariance matrix S_w according to Raducanu and Vitriá (2008) is expressed as:

$$S_w = \sum_{i=1}^{C_N} \sum_{j \in C_i} (m_j - \mu_{C_i}) (m_j - \mu_{C_i})^T, \quad (3.1)$$

where, μ_{C_i} and m_j are the mean vector and sample of class C_i respectively. The between class covariance matrix S_b [18] is expressed as follows:

$$S_b = \sum_{i=1}^{C_N} \sum_{j=1, j \neq i}^{C_N} \sum_{q=1}^{\omega_{C_i}} W_{ijq} (m_q^i - \mu_{NN}(m_q^i, C_j)) (m_q^i - \mu_{NN}(m_q^i, C_j))^T, \quad (3.2)$$

where, ω_{C_i} is the number of samples in class C_i . m_q^i is the q^{th} sample of class i . In S_b , $\mu_{NN}(m_q^i, C_j)$ is local k -NN mean defined as:

$$\mu_{NN}(m_q^i, C_j) = \frac{1}{k} \sum_{t=1}^k NN_t(m_q^i, C_j), \quad (3.3)$$

where, $NN_t(m_q^i, C_j)$ is t^{th} nearest neighbour from vector m_q^i to class C_j . W_{ijq} is a weighting function used in the between class covariance matrix denoted as:

$$W_{ijq} = \frac{\min \{d^\alpha(m_q^i, NN_t(m_q^i, C_i)) (m_q^i, NN_t(m_q^i, C_j))\}}{d^\alpha(m_q^i, NN_t(m_q^i, C_i)) + (m_q^i, NN_t(m_q^i, C_j))}, \quad (3.4)$$

where, α denotes control parameter for sample weights which can be selected between zero and infinity. The class separability according to Raducanu and Vitriá (2008) can be expressed as:

$$\Omega_{NDA} = tr((S_w)^{-1} \cdot (S_b)). \quad (3.5)$$

3.3 Incremental Nonparametric Discriminant Analysis (IncNDA)

According to Raducanu and Vitriá (2008), a situation where the new instances are coming in, the IncNDA can be defined as:

$$\Omega_{IncNDA} = tr \left(\left(S_w' \right)^{-1} \cdot \left(S_b' \right) \right), \quad (3.6)$$

where, S_w' , and S_b' are the updated within class and between class covariance matrix. For more detail on Ω_{IncNDA} refer to Raducanu and Vitriá (2008). Let the incoming data y belong to one of the existing classes C_L (i.e. y^{C_L}). The updated between class S_b' and within class S_w' covariance matrix are defined as follows:

$$S_b' = S_b - S_b^{\text{in}}(C_L) + S_b^{\text{in}}(C_{L'}) + S_b^{\text{out}}(y^{C_L}), \quad (3.7)$$

$$S_w' = \sum_{j=1, j \neq L}^{C_N} S_w(C_j) + S_w(C_{L'}), \quad (3.8)$$

where, the covariance matrices $S_b^{\text{in}}(C_{L'})$, $S_b^{\text{out}}(y^{C_L})$ and $S_w(C_{L'})$ are expressed as :

$$S_b^{\text{in}}(C_L) = \sum_{j=1, j \neq L}^{C_N} \sum_{i=1}^{n_{C_j}} W_{ijL} \left(m_i^j - \mu_{NN}(m_i^j, C_L) \right) \left(m_i^j - \mu_{NN}(m_i^j, C_L) \right)^T, \quad (3.9)$$

$$S_b^{\text{out}}(y^{C_L}) = \sum_{j=1, j \neq L}^{C_N} (y^{C_L} - \mu_{NN}(y^{C_L}, C_j)) (y^{C_L} - \mu_{NN}(y^{C_L}, C_j))^T, \quad (3.10)$$

$$S_w(C_{L'}) = S_w(C_j) + \frac{\omega_{C_L}}{\omega_{C_L} + 1} (y - \mu_{C_L})(y - \mu_{C_L})^T. \quad (3.11)$$

The covariance matrix between the existing class and the class about to be updated is denoted by $S_b^{\text{in}}(C_L)$. The covariance matrix between the existing class and the updated class $C_{L'}$ is denoted by $S_b^{\text{in}}(C_{L'})$ and $S_w(C_{L'})$ signifies the updated within

class covariance matrix. ω_{C_L} is the number of samples in class C_L . L denotes the new instances belonging to new class C_L .

3.4 Selective Sampling Criterion for Incremental Learning

As mentioned previously, most of the existing active learning models are implemented in supervised or semi-supervised learning setting. In our research we aim to develop an unsupervised active learning model. Therefore we have introduced two different criterions which would eliminate the oracle/supervisor entity and making our model an unsupervised active learner.

In this criterion if the classification accuracy (obtained through SVM) of the current chunk of incoming data is less than the previous classification accuracy value, then that data chunk is assumed to be non-informative, and is not used for incremental learning.

3.4.1 Classification Accuracy Criterion (CAC)

Let Classification Accuracy Criterion (CAC) (λ) for each learning stage be such that $\lambda > \theta$, where θ is the threshold value (Classification accuracy) obtained from the previous stage of incremental learning. If $\lambda > \theta$, then the given data chunk is informative, hence selected for assimilation in incremental learning.

Let Y_0 be the original data and Y_t represent the incoming data chunk where $t = 1, 2, 3, \dots, N$.

If $\lambda_t > \theta$ then

$$Y'_t = Y_t + \sum_{i=1}^{t-1} (Y_i) \text{ where } \begin{matrix} t = 0 \\ t \neq 0 \end{matrix}. \quad (3.12)$$

Algorithm: IncNDA based Active Learning Model with CAC

Input:

Y_0 – training data.

Y_t - incoming data chunk where $t = 1, 2, 3... , N$.

θ - Threshold Value of Y_0 .

λ - criterion

Calculate NDA + SVM of Y_0 .

if $\lambda_t < \theta$ **then**

process next incoming data chunk Y_t

else

Select incoming data chunk Y_t

for $i = 1$ to N

$Y_t' = f(Y_t) + Y_i$ when $t = 0$ and $t \neq 0$.

Calculate Ω_{IncNDA} and SVM of Y_t .

Obtain λ from SVM of Y_t .

end for

end if

Output:

Updated Criterion λ_t

Updated Eigen Space of IncNDA model Ω_{IncNDA}

Tuned SVM Classifier Ω_{SVM}

3.4.2 Boundary Class Information Criterion (BCIC)

We have submitted to “*The Eighth International Conference on Information and Management Sciences*” one of our recent studies on active learning which used the discrimination residue based criterion mentioned in Pang et al. (2009).

The idea for BCIC has been adapted from Eq. (3.4), where weighting function is used to emphasize the boundary class information. BCIC measures the mean weighted distance of within and between class vectors. Since each vector of class C_i points towards the local mean distance of class C_j the scatter matrix of these vectors show the subspace in which the class boundary is embedded. According to Raducanu and Vitriá (2008), since sample weights value tend to 0.5 on class boundary and drops to zero as we move away from the classification boundary. Therefore BCIC reveals if the incoming data has important classification boundary information. BCIC criterion (w_{ij}) is defined as:

$$w_{ij} = \frac{d^\alpha(m_q^i, NN_t(m_q^i, C_i))}{\frac{1}{n} \sum_{i=1}^n d^\alpha(m_q^i, NN_t(m_q^i, C_j))}. \quad (3.13)$$

If $w_{ij}' < \theta$, where θ is the threshold value, then Eq. (3.12).

Discrimination Residue Ratio based Criterion

For active learning, we consider here an active learning way (aIncNDA) to empower the IncNDA with the ability of detecting the discriminative interestingness of data; before it is delivered for IncNDA learning. That is, the above IncNDA can be renovated to conduct incremental learning in an active learning way as:

$$\Omega(t+1) = \begin{cases} F_c(\Omega(t), y) & \text{if } L(t) > \xi \\ \Omega(t) & \text{otherwise.} \end{cases}, \quad (3.14)$$

where, only discriminative instances are delivered for IncNDA learning. ξ is the threshold identifying discriminative criterion of NDA. The smaller ξ leads to the bigger number of instances learned by IncNDA.

Recall that the nature of NDA learning lies at the discriminability difference between the NDA transformed space and the original space. Straightforwardly, $L(t)$ can be represented as a type of mathematical residue that reflects the discriminability difference between the NDA transformed space and the original space.

Given one new instance is presented at any given time, the discriminability difference between the NDA transformed space and the original space of the IncNDA at time t by a classification performance evaluation as:

$$L(t) = Ad(t) - Ao(t), \quad (3.15)$$

where, $Ad(.)$ is the classification accuracy on discriminant eigenspace, and $Ao(.)$ is the accuracy on original space. It could be any type of classification performance evaluation by any classifier.

However, such performance-based residue calculation involves a serious problem.

That is, the $L(t)$ is highly classifier dependent. For example, suppose a k-NN method is used for performance evaluation $Ad(.)$ and $Ao(.)$, then the selected instances for incremental learning is meaningful only for k-NN classification and the category of prototype-based methods, but not for the classification using any other methods such as hyperplane-based support vector machines (SVM) and decision-tree based C4.5.

Discrimination Residue Ratio

The idea of discrimination residue ratio is adapted from the weighting function (i.e. Eq. 3.4) used in NDA, where $NN_k(x_i, C_i)$ and $NN_k(x_i, C_j)$ emphasize local within class distances and local between class distances. As we know, the principle of NDA, similar to LDA, seeks simultaneously minimizing within class distances and maximizing between class distances. The difference between NDA and LDA is that LDA is global model, whereas NDA focus on local instances distribution.

Given M new instances $Y = \{y_1, y_2, \dots, y_M\}$ presented as one chunk at time t , for each instance $y_i \in Y$, we can quickly estimate the within-class residue to the class mean vector μ_{C_i} :

$$\|NN_k(y_i, C_i) - \mu_{C_i}\|, \quad (3.16)$$

also the between-class residue to any other class mean vector μ_{C_j} , $j = 1, \dots, C_N, j \neq i$:

$$\|NN_k(y_i, C_j) - \mu_{C_j}\|. \quad (3.17)$$

Thus, the contribution of incoming instance y_i to the NDA fundamental maximum $tr(S_w.S_b)$ criterion can be estimated as the following discrimination residue ratio of within-class to between-class scatter estimates:

$$v(y_i) = \frac{\|NN_k(y_i, C_i) - \mu_{C_i}\|}{\left\| \frac{1}{C_N-1} \sum_{j=1, j \neq i}^{C_N} NN_k(y_i, C_j) - \mu_{C_j} \right\|}. \quad (3.18)$$

If $v(y_i) > 1$, then the contribution of y_i to NDA discrimination is positive, otherwise it is negative.

However, it is noticeable that the above discrimination residue ratio varies in practice largely depending on the individual dataset. Thus, it is hard for us to determine a suitable threshold value for a given dataset. To overcome this difficulty, we compute

the discrimination residue ratio for every instance of the Y , then the above $v(y_i)$ can be normalized as:

$$v_{y_i} = \frac{v - \bar{v}}{\sqrt{\frac{1}{M} \sum_{m=1}^M (v_m - \bar{v})^2}}, \quad (3.19)$$

where, $\bar{v} = \frac{1}{M} \sum_{m=1}^M v_m$ is the chunk mean discrimination residue ratio. Thus, $L(t)$ in Eq.(3.20) can be implemented by v_{y_i} as a chunk data filter.

$$\Omega' = \begin{cases} F_c(\Omega, y) & \text{if } v(y) > \xi \\ \Omega & \text{otherwise.} \end{cases} \quad (3.20)$$

In the above sections, we have covered various criteria and discriminative methods used in our novel active learning framework. In the following section our active learning framework has been presented.

3.5 Active Learning Framework

Figure 3.1 shows how the proposed novel active learning method works and how the base classifier and sampling algorithm are going to be integrated into the active learner to improve its overall efficiency. In the proposed model an incremental NDA has been implemented as an adaptive sampling method along with SVM and k-NN which is the base classifier. The rationale for using incremental learning is to reduce the time and computational cost of having to evaluate new streaming data from scratch.

As seen in Fig.3.1, the initial data will be evaluated based on the NDA's analysis to form the initial discrimination rule and at the same time the base classifier will be trained on it to form an initial classification boundary. When new streaming data is available, selective sampling of the data will be done based on a criterion. If the new data is found to be relevant then it is added to NDA's incremental learning memory otherwise NDA continues looking for relevant data in the next incoming network data. Moreover, if relevant data is found by the criterion, then NDA increments the data which is further tested using (SVM/k-NN) classifier. Active learning model continued the same process each time new streaming data is available.

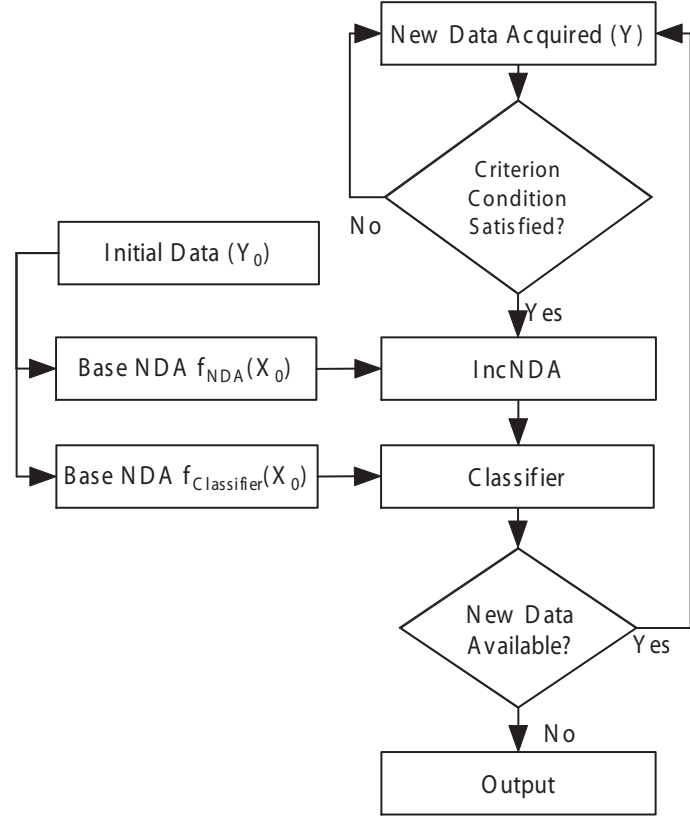


Figure 3.1: Active mode IncNDA learning model (aIncNDA).

3.5.1 Active mode IncNDA model (aIncNDA)

Let X_0 denote the initial data. Nonparametric discriminant analysis is conducted on X_0 to build the initial discrimination rule represented by $f_{\text{NDA}}(X_0)$. Also the classifier is trained on X_0 to build an initial classification boundary $f_{\text{Classifier}}(X_0)$ for decision making.

Let Y denote the newly acquired dataset from network. On the data Y , the following

incremental NDA is carried out to perform discrimination measure:

$$f'_{\text{NDA}} = \Omega_{\text{IncNDA}}(f_{\text{NDA}}, Y), \quad (3.21)$$

where, Ω_{IncNDA} represents the Incremental NDA model with the newly acquired relevant data Y . Based on Eq.(3.21), we evaluate the discrimination value of data Y by computing the difference between f'_{NDA} and f_{NDA} .

If

$$\|f'_{\text{NDA}} - f_{\text{NDA}}\| < \theta, \quad (3.22)$$

where, θ is the threshold value determined after model generation, then Y is not relevant data and NDA continues looking for relevant data by analyzing the next chunk from the network.

On the other hand, if

$$\|f'_{\text{NDA}} - f_{\text{NDA}}\| > \theta, \quad (3.23)$$

then the data Y is relevant and thus is used to Increment NDA as Eq.(3.21). Furthermore, classifiers (such as SVM/k-NN) are used to perform the classification on data Y .

Algorithm: aIncNDA learning algorithm

Input:

Y_0 – training data.

Y_t - incoming data chunk where $t = 1, 2, 3... , N$.

λ - criterion

Calculate NDA of Y_0 .

if λ condition satisfied **then**

select incoming data chunk Y_t

else

process to next incoming data chunk Y_t

for $i = 1$ to N

Calculate Ω_{IncNDA} of Y_t .

Obtain updated λ from Y_t .

end for

end if

Output:

Updated Criterion λ_t

Updated Eigen Space of IncNDA model Ω_{IncNDA}

Tuned Classifier $\Omega_{\text{Classifier}}$

The above active learning steps continue until all the data samples have been tested. As a result, the base classifier (such as SVM or k-NN) is tuned to an optimum level, giving a continuously improved SVM or k-NN on classification.

3.6 Summary

In this chapter, we have discussed the discriminative techniques used in our framework, introduced two criteria that are used for selective sampling of informative/relevant instances. We have described our novel active learning model and its mechanism. In the next chapter, the system configuration used for implementing our methods is presented along with the experiments and analysis on the benchmark UCI datasets.

Chapter 4

Discrimination Experiments on the Benchmark Datasets

A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.

- Albert Einstein

4.1 Introduction

In this chapter, we begin by specifying software and hardware configuration used for implementing the novel framework and method for active learning, followed by a demonstration of incremental learnings significance and characteristics, such as its ability to preserve the information obtained from previous data chunks, therefore eliminating the need for storing previous data and recalculation of scatter matrix. We have then presented the quantitative comparison of IncNDA with Active Learning model. The active learning models were tested using different criterions.

4.2 System Configuration

4.2.1 Software Configuration

- **Operating System**

Microsoft Windows XP, Version 2002 operating system has been used for implementing this project. It is the standard operating system available in *Knowledge Engineering and Discovery Research Institute* (KEDRI) facility. It is the first consumer-oriented operating system produced by Microsoft to be built on the Windows NT kernel and architecture.

- **Programming Language**

MATLAB Version 7.5.0.342 (R2007b) has been used for implementing the novel framework and method for active learning. The novel framework and method for *Active Learning* have been entirely programmed in MATLAB. MATLAB is a fourth generation programming language which provides a numerical computing environment. MATLAB was chosen as a programming platform as it allows easy matrix manipulation, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs in other languages.

4.2.2 Hardware Configuration

An Intel Core 2 CPU, 1.86 GHz, with 1.99 GB of RAM was used. This is a standard hardware configuration provided by KEDRI.

To demonstrate the workings and efficiency of (NDA's) incremental learning, a small experiment on the benchmark dataset from UCI archives has been conducted, and its results can be seen in table 4.1 and figure 4.1.

It can be observed that on most of the dataset; IncNDA's classification performance is on par with other discriminative methods such as LDA and SVM.

Table 4.1: A comparison between SVM, LDA and IncNDA in terms of classification accuracy and confidence interval.

Dataset	(Batch) SVM	(Batch) LDA	IncNDA
Iris	94.00% +/- 10.63	95.33% +/- 5.49	96.00% +/- 4.66
Wisconsin	95.25% +/- 2.63	95.15% +/- 3.97	95.06% +/- 4.82
Heart	58.97% +/- 4.80	60.25% +/- 9.35	60.60% +/- 6.18
Glass	60.67% +/- 13.97	35.61% +/- 12.48	58.83% +/- 9.82
Ionosphere	88.29% +/- 4.56	90.29% +/- 5.08	92.43% +/- 6.48
LiverDisorder	68.71% +/- 8.16	69.82% +/- 8.70	69.53% +/- 7.89

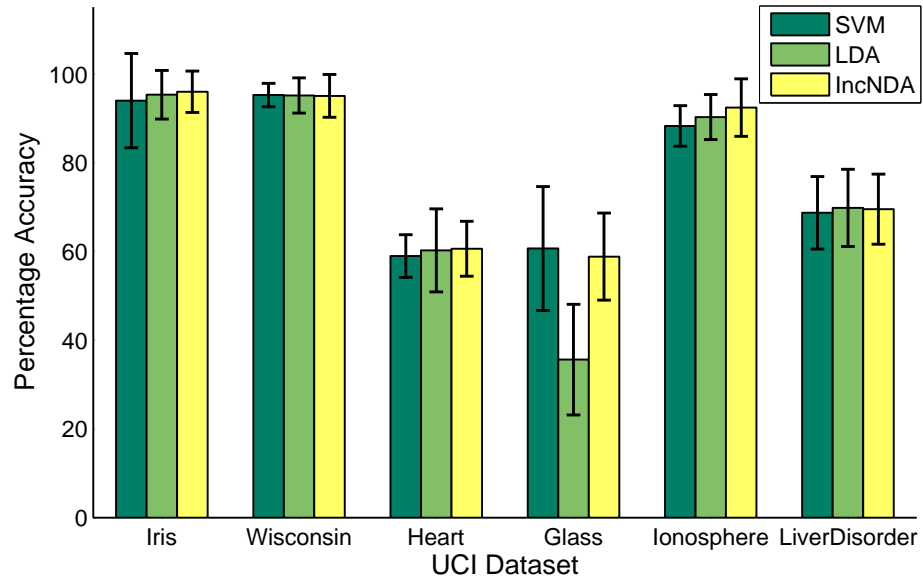


Figure 4.1: Performance evaluation of Classifiers.

Incremental Chunk Learning process:

The table 4.2 shows the stages of learning process after each update of the initial NDA eigenspace. The graph falls some time because the percentage of total data been calculated is relative to the number of classes obtained at that given stage.

The updated data is provided in a total of 10 stages. At each stage 10 percent of the total data is present (randomly) such that, the samples are not recurring. Therefore, each update has unique non-overlapped training samples.

Table 4.2: Incremental learning performed by NDA at different stages for Iris dataset. At every learning stage, 10% of the total data is provided. The percentage accuracy was calculated using leave-one-out cross validation.

Iris Dataset	
Learning Stage	Classification Accuracy
Stage 1	80.00 +/- 34.96%
Stage 2	85.00 +/- 24.15%
Stage 3	80.00 +/- 23.31%
Stage 4	91.00 +/- 11.74%
Stage 5	91.67 +/- 11.79%
Stage 6	95.00 +/- 8.74%
Stage 7	94.44 +/- 7.86%
Stage 8	95.27 +/- 4.99%
Stage 9	96.67 +/- 5.83%
Stage 10	96.26 +/- 3.94%

Incremental Eigen Space Model

In this method the eigenspace is updated incrementally with a new single training sample at a time. At each stage as new updates are introduced all the previous stage updates are added to the initial training set making it a one - pass incremental learning method.

This approach creates a problem since it requires all the updates to be added to the training set; thereby, requiring increasing memory space after each update, but due to the availability of previous updates an accurate eigenspace is obtained.

In the following, we have presented experiment and analysis of aIncNDA method using two criteria on benchmark datasets.

4.3 Experiment 1: CAC Criterion

4.3.1 Experimental Setup

Data Chunks (Learning Stages): As a general framework for our experimentation, we have randomized and split the dataset into 10 data chunks (Y_t), therefore each data chunk represents 10 percent of the whole dataset. The initial data chunk Y_0 is used for training and the rest of the data chunks are used for active incremental learning.

For Discrimination based Active Learning the incoming data chunk $Y_t\{t \neq 0\}$ is projected into eigenspace for pattern classification. Selective sampling is performed using k -Nearest Neighbor classifier, where the samples k closest neighbors are selected.

Missing Values: In datasets having missing values, the missing values have been replaced by the mean value of that attribute's column. Also, the datasets have been applied with a filtering algorithm which removes redundant instances if any.

Parameter Settings: Throughout our experiment, SVM and LDA from NeuCom Student v0.917 was used. For SVM, we have implemented a 10-fold cross validation technique without normalization and feature selection. A polynomial SVM kernel with inductive training method was used. SVMs degree constraint was selected as 1. For NDA the weighting function was set to 0.5 with nearest neighbor as 2, 5 and 7

UCI Database

The performance evaluation was carried out on benchmark datasets obtained from UCI machine learning repository. A total of 13 datasets are used for performance evaluation of aIncNDA versus NDA, where each dataset differs in terms of number of classes, features, class distribution etc. The description of the UCI datasets used in this performance evaluation is given in Table 4.3.

Table 4.3: Summary of evaluated UCI datasets.

Dataset Index	Dataset Name	Classes	Attributes	Instances
D1	Wisconsin	2	11	699
D2	Ionosphere	2	34	351
D3	Liver Disorder	2	6	345
D4	German Credit Data	2	20	1000
D5	Pima Indians Diabetes	2	8	768
D6	Hepatitis	2	19	155
D7	Iris	3	4	150
D8	Wine	3	13	178
D9	Heart	5	14	297
D10	Glass	7	10	214
D11	Ecoli	8	8	336
D12	Vowels	11	10	528
D13	Face	169	100	845

4.3.2 Results

Table 4.4: Comparison between aIncNDA (Active) and IncNDA learning in terms of classification accuracy. Keys: DI = Dataset Index, SS = Samples selected by aIncNDA for learning, NN = Nearest Neighbour.

DI	2 NN			5 NN			7 NN		
	IncNDA	aIncNDA		IncNDA	aIncNDA		IncNDA	aIncNDA	
		SS	Active		SS	Active		SS	Active
D1	91.14%	33%	90.92%	91.57%	50%	92.00%	91.57%	63%	95.24%
D2	80.00%	33%	74.28%	73.42%	47%	84.00%	84.00%	66%	83.71%
D3	62.17%	45%	62.17%	68.32%	38%	72.14%	70.96%	65%	71.26%
D4	68.80%	57%	69.90%	67.80%	53%	68.70%	71.60%	61%	69.80%
D5	69.55%	25%	70.01%	75.03%	38%	73.82%	74.27%	42%	75.79%
D6	81.93%	36%	79.35%	81.93%	61%	83.22%	80.00%	59%	85.80%
D7	95.30%	65%	95.97%	95.97%	66%	96.64%	95.97%	60%	95.97%
D8	85.95%	84%	93.82%	91.01%	83%	92.69%	56.17%	69%	92.13%
D9	52.86%	43%	54.54%	53.19%	36%	54.54%	52.18%	65%	54.88%
D10	56.33%	66%	54.46%	59.62%	58%	58.68%	61.91%	51%	61.97%
D11	77.84%	68%	78.14%	81.43%	28%	81.43%	85.32%	50%	84.43%
D12	96.53%	83%	95.67%	83.98%	69%	84.19%	77.05%	61%	83.54%
D13	85.90%	39%	86.37%	75.47%	66%	76.18%	69.78%	83%	70.85%

4.3.3 Discussion

Table 4.4 shows the final percentage accuracy for three different nearest neighbour setting. The accuracy is different for each nearest neighbour setting due to the different number of instances, classes and their data distribution. The updated data is provided in a total of 10 stages. In each stage 10 percent of the total data is present (randomly) such that the samples are not recurring therefore each update has unique non-overlapped training samples therefore making it a one - pass incremental learning method. Comparing the performance of active learning with criterion is close to and in some datasets better than IncNDA. Furthermore, the classification accuracy achieved is justifiable since the number of samples selected through selective sampling is less than the total number of samples. This shows that the selective sampling criterion performs well on most of the datasets. However, there is a need for a more robust criteria therefore in future works experimentation with different criterion. Also, a correlative or self organizing map (SOM) based method will be explored as an alternative to the current euclidean based nearest neighbour selection for within and between-class in NDA learning.

4.4 Experiment 2: BCIC Criterion

We have submitted for publication one of our recent studies on active learning which used the discrimination residue based criterion mentioned in Pang et al. (2009).

In this section, we have examined the efficiency and accuracy of the proposed aIncNDA method, and compared to IncNDA. Particularly, we investigate the relationship between 1) the discriminability and number of instances, 2) the redundancy and number of instances. To experiment on data with different discriminative characterization, we used datasets from two database resources. One resource is from UCI Machine Learning Repository (Hettich & Bay, 1999), where we selected 8 datasets that have different application backgrounds and the features 100% of continuous/integer values and no missing value. The other resource is the MPEG-7 face database (Kim, Kim & Lee, 2003), which consists of pose and light two subsets, total 1355 face images of 271 persons, 5 different face images per person and each face image has the size of 56 x 46.

4.4.1 Experimental Setup

To implement the proposed aIncNDA for incremental learning using this criterion, we randomly select 10% of data from each dataset, for initial batch NDA training, and divide the remaining data into 10 random chunks for incremental learning test. We collect every instance learned by aIncNDA, and evaluate the performance of aIncNDA and IncNDA on discrimination contribution at every learning stage. For performance evaluation, we compared the eigenspace from the proposed aIncNDA with the eigenspace from IncNDA by a leave-one-out k-NN ($k=1$) classification over all data presented by current learning stage. Note that we use the term learning stage instead of the usual time scale since the events of data arriving in the above incremental learning may not happen in a regular time interval. Here, the number of learning stages is equivalent to the number of instances that have been learned by incremental models.

4.4.2 Results

In the experiment, parameter ξ is relevant to the number of curiosity instances and the discriminability of the resulting NDA. For each experiments, we fixed ξ by the rule that the instances are significantly selected with, at most, minor sacrifices in discriminability.

Synthetic Data

We first experimented the proposed aIncNDA with a synthetic data set that has 3 classes 475 instances. The data distribution is a mixture of several 2D ($[X1 \ X2]$) Gaussian distributions as shown in Fig. 4.2. Fig. 4.3 gives the distribution of the 257 informative instances learned by aIncNDA. As compared to the data distribution of the entire 475 instances, the discriminative representativeness of the selected instances by aIncNDA is clear because those 257 instances includes all critical instances for class distinction, such as instances involving class-mixture, and major representative instances of the independent class.

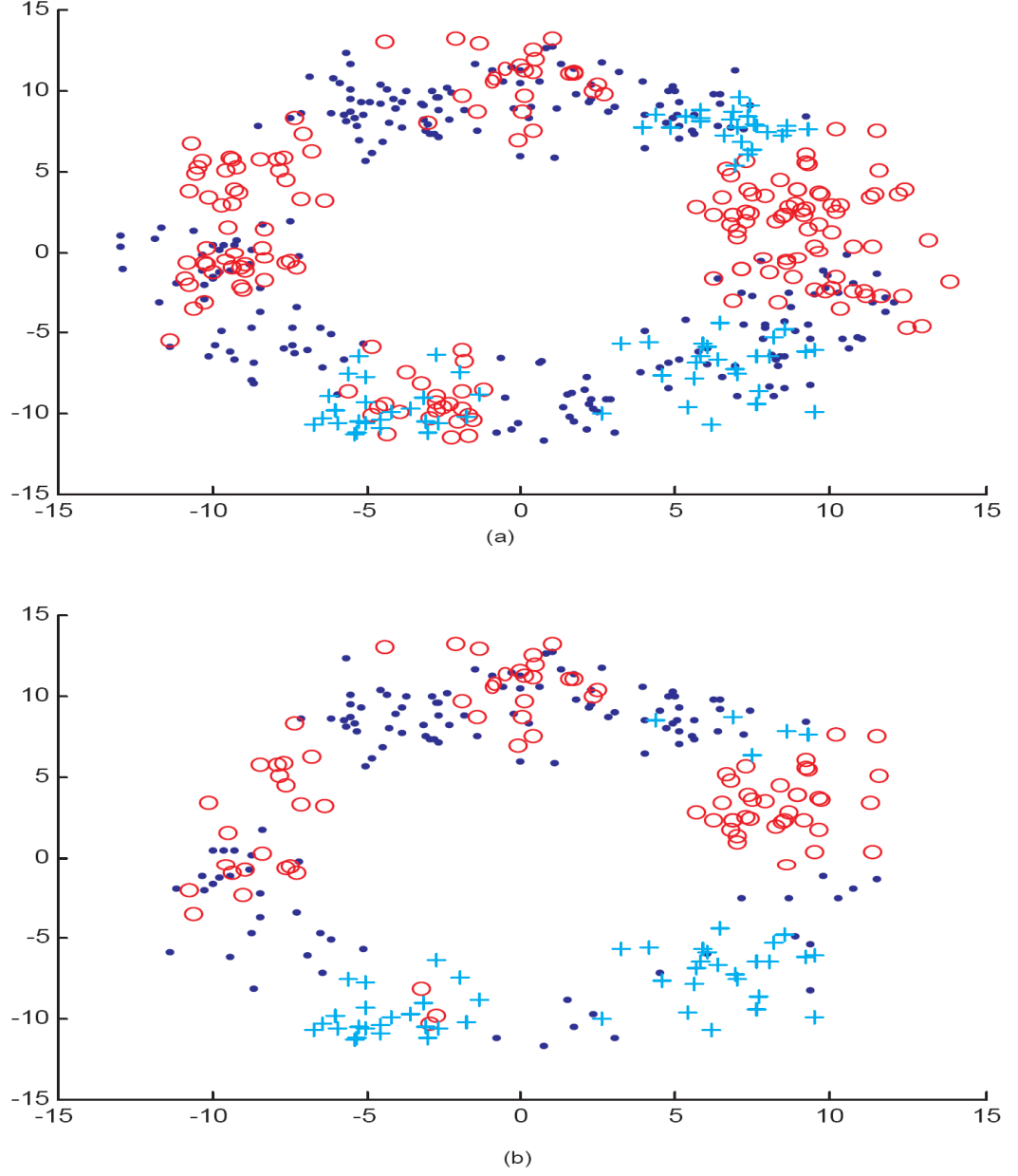


Figure 4.2: The comparison of data distribution between the synthetic dataset and selected curiosity instances by proposed aIncNDA learning method. (a) The data distribution of the entire dataset; and (b) The data distribution of selected instance by aIncNDA.

Fig. 4.3 illustrates the whole procedure of incremental learning with a comparison to IncNDA, where the horizontal and vertical axis represent the incremental stage and the classification accuracy from k-NN ($k=1$). As seen from the figure, the proposed aIncNDA and IncNDA is compared on the classification error at every in-

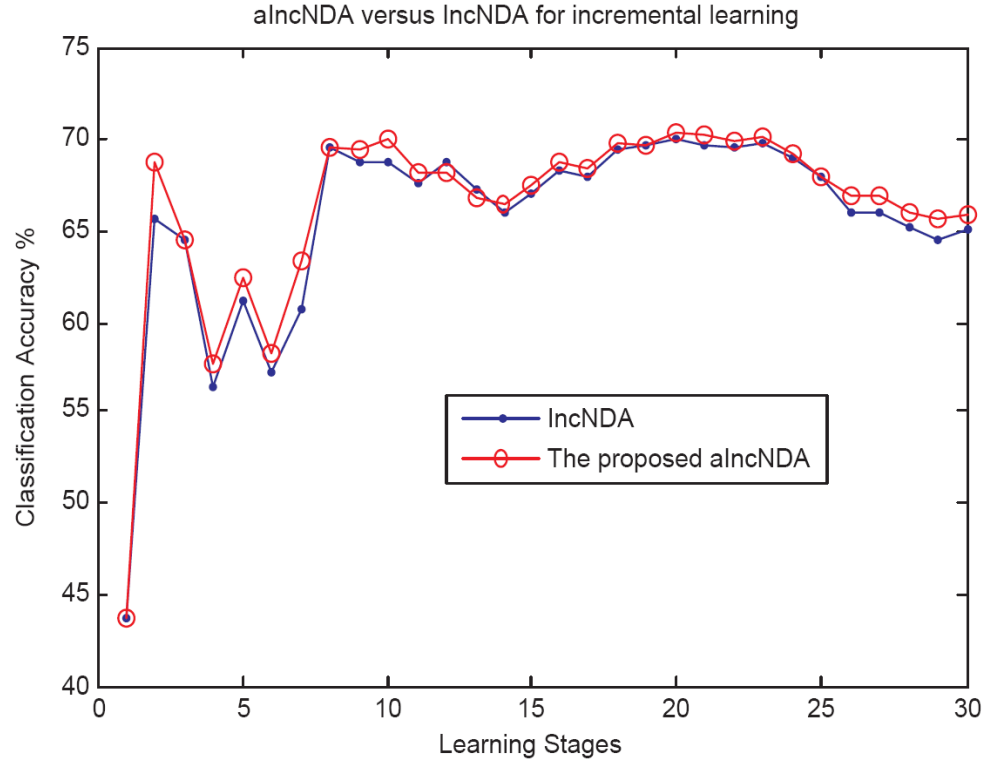


Figure 4.3: The comparison of aIncNDA and IncNDA on the performance of incremental learning.

cremental learning step. The classification accuracy difference between two methods is $+0.842105$, which indicates that the proposed aIncNDA achieves better learning effectiveness of the original IncNDA, although aIncNDA learns only 54.10% of total 475 instances.

UCI Datasets Table 4.5 gives an comparison of aIncNDA versus IncNDA on the incremental learning of 8 UCI datasets. In the table, ξ is fixed for each dataset by the rule described above, the number of instances and the percentage to the number of all instances is denoted as ‘No.Instances (rate%)’, the classification accuracy as ‘Acc’, and the discriminability difference (denoted as ‘Diff.’) is calculated as the proposed aIncNDA minus IncNDA in terms of the k-NN leave one out classification performance at the final learning stage.

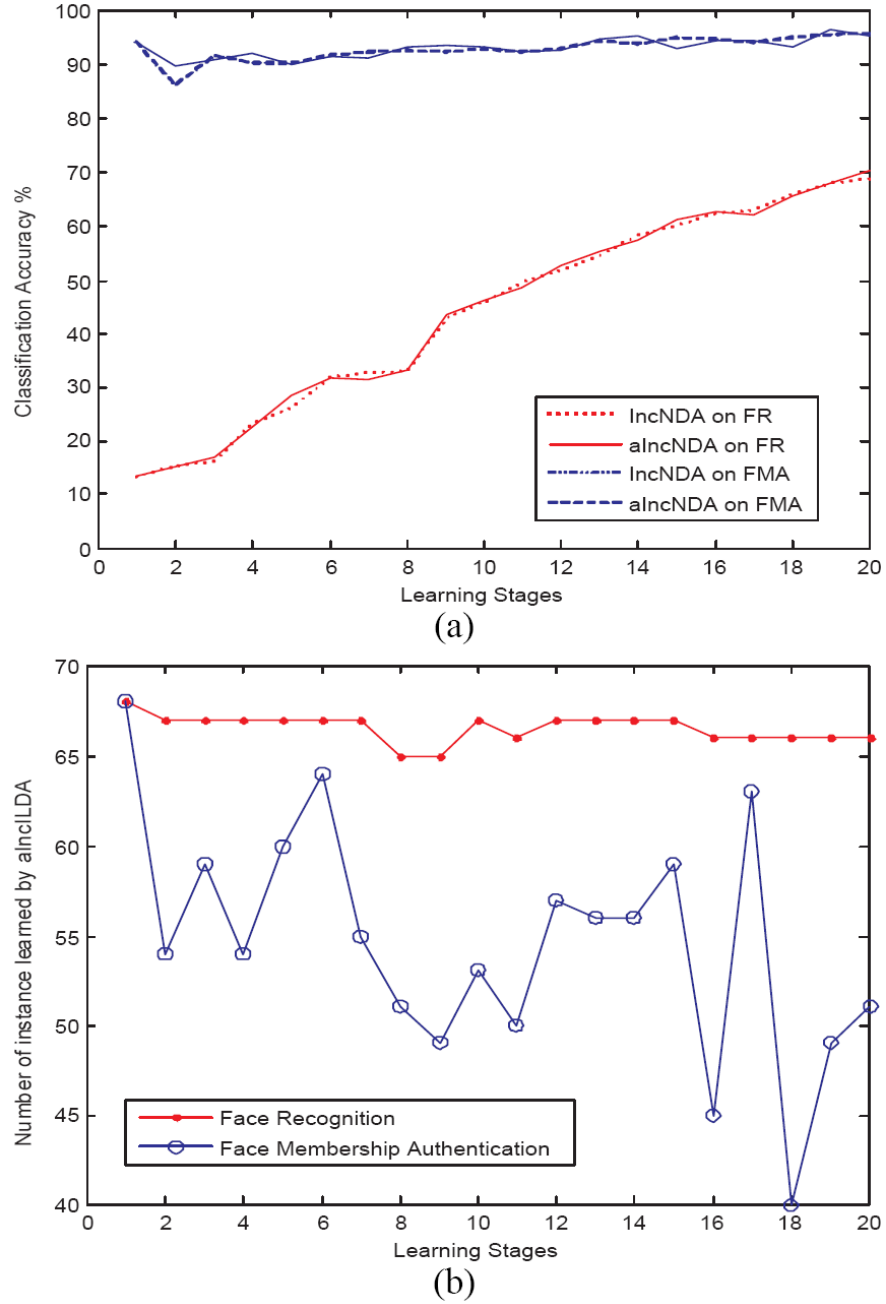


Figure 4.4: Comparison of aIncNDA and IncNDA on FR and FMA, (a) the performance of aIncNDA versus IncNDA on incremental learning; (b) the number of learned instances by aIncNDA at every learning stage. (using Discriminant Ratio based criterion)

As seen in table 4.5, the proposed aIncNDA method, ignores 4.70% - 88.90% instances of the whole dataset, constructs discriminant eigenspaces on the remaining 11.10% -

Table 4.5: Comparison of aIncNDA versus IncNDA on Incremental Learning over 8 UCI benchmark datasets.

	aIncNDA			IncNDA		
Dataset	ξ	No.Instances(rate%)	Acc.%	No.Instances	Acc.%	Diff.%
Iris	0.75%	56 (37.3)%	94.5%	150	92.0%	+2.5%
Liver-disorder	0.8%	51 (22.2)%	63.3%	345	62.4%	+0.9%
Vehicle	3.0e-3%	251 (29.7)%	77.6%	846	75.4%	+2.2%
Glass	0.98%	50 (23.4)%	60.1%	214	52.5%	+7.6%
Wine	0.95%	162 (92.7)%	77.6%	846	75.4%	+2.2%
Wisconsin	0.95%	443 (95.7)%	84.3%	463	89.7%	+1.1%
Ionosphere	0.7%	291 (83.1)%	76.2%	350	76.1%	+0.1%
Heart	0.65%	33 (11.1)%	53.2%	297	52.3%	+0.9%

95.30% selected instances. However, the discriminability of the obtained eigenspace from composed instance subset, compared to the eigenspace from all instances (using IncNDA), has no decrease; reversely, in most cases it has a slight increase. This suggests that the proposed active IncNDA learning is valid, and the selected instances by aIncNDA have the expected discriminative representativeness.

4.4.3 Discussion

To test the performance of the proposed method under different level of discriminative redundancy, we carried out face recognition (FR) and face membership authentication (FMA) experiments using the same face database described above. FMA is to distinguish the membership class (cls.1) from the non-membership class (cls. 2) in a total group through a binary class classification. FMA involves more discriminative redundancy than face recognition problem, because the size of membership in FMA is often smaller than that of nonmembership, which indicates that not every instance is discriminatively important for FMA. Over the 271 persons 1355 faces data, we conducted FR and FMA, respectively. For the FMA experiment, we set the membership size as 71 (cls. 1/cls. 2 is 71/200) without loss of generality. Thus, we compared the proposed aIncNDA with the IncNDA on incremental learning of 271 classes (i.e. FR) and 2 classes (i.e. FMA) data, respectively. Fig. 4.4(a) shows the comparison of NDA discriminability between the proposed aIncNDA and the IncLDA for both FR and FMA experiments, and Fig. 4.4(b) reports corresponding

the number of instances learned by aIncNDA. As seen in Fig. 4.4(a), the proposed aIncNDA learns NDA for FR on 1331 of total 1355 instances, only 24 instances are found redundant. Whereas for FMA, aIncNDA learns 1093 of 1355 which means only about 20.0% of all 1355 instances are reduced. However, the performance of the proposed aIncNDA for both FR and FMA as given in Fig. 4.4(a) outperforms the performance of the IncNDA in most cases, on all 1355 instances. This indicates that the proposed aIncNDA is able to adapt itself automatically to data with discriminative redundancy, and select a suitable number of instance to build an correct NDA model. This characteristic/property can also be observed in Fig. 4.4(b), where aIncNDA actively selects a particular number of discriminative instances for incremental learning.

Over the datasets from different resources, the proposed aIncNDA learning method is evaluated on: (1) aIncNDA versus IncNDA, and (2) performance under different level redundancy, where face recognition and face membership authentication are studied, respectively. The experimental results demonstrate that the proposed aIncNDA learning helps more efficient NDA learning with fewer instances, without any performance reduction. One limitation of the proposed method is that, the data processing in aIncNDA is not one-pass as the original IncNDA retains raw data at every step of incremental learning.

A method based on passive learning proves to be inadequate in real world application. To overcome this limitation, we have developed active mode incremental NDA which performs adaptive discriminant selection of instances for incremental NDA learning. Performance evaluation carried out on benchmark UCI datasets shows that Active Mode Incremental NDA performs on par and in many cases better than incremental NDA with fewer instances. Given the nature of network data which is large, streaming, and constantly changing, we believe that our method can find practical application in the field of internet security.

Future Works As future work, the presented methods application in intrusion detection system will be explored along with added enhancements to the selective sampling criterion. Also, the use of incremental classifier will be researched to serve as an extension to our present model which will eliminate the need for retraining, further enhancing the processing speed while been computationally efficient.

4.5 Summary

In this chapter we have presented experiments and discussion using the two mentioned criteria in our active learning framework. The experiments in this chapter have been performed on benchmark datasets from UCI archives which are available online. The experiments performed using the discriminative ratio criteria have been submitted to ‘*The Eighth International Conference on Information and Management Sciences*’ (Pang et al., 2009), which has also been referred to as aIncNDA in this chapter. The next chapter will present the application of our novel active learning technique for image recognition and retrieval task.

In this chapter we have underlined the system configuration used for implementing and materializing our research. The next chapter presents the experiments and analysis which have been performed on benchmark dataset using two different criteria independently.

Chapter 5

Multi-example Image Retrieval Applications

This chapter presents a novel application of multi-example image retrieval based on active mode incremental nonparametric discriminant analysis learning (MeIR).

Traditional methods conduct query using only one image as the template for similarity comparison while retrieving. Alternatively in our method, the template is replaced by a sort of discriminative differences amongst multiple of example images. The discriminative differences extracted out by NDA from the given set of example images is used along with correlation based similarity metrics. The extracted image samples are incrementally and iteratively learned in order to obtain next correlated images from the image dataset.

Though there have been advancements and some success in retrieval of textual information, very little has been achieved in the image retrieval domain (Datta, Joshi, Li & Wang, 2008). Majority of the techniques developed for image retrieval permit only single example as a query. In a situation where the user needs to query multiple images, this causes an inconvenience, since the user can only input one image as a query at a given time.

Therefore, in this chapter, we have applied the aIncNDA for image retrieval that allows multiple examples of images for querying. The advantage of our proposed method is three folds, firstly, MeIR retrieves the images by exploiting the discriminative features present in the multi-example query. Secondly, it addresses the

usability issues present in single image query based retrieval system. Thirdly, the adaptive and incremental learning nature of the proposed method allows us to efficiently recognize, acquire/retrieve a more refined image search result, with the least number of images for training.

5.1 Introduction

The explosive growth and accessibility to technologies such as digital cameras and the World Wide Web, has allowed the public access to large amounts of data. Retrieval of the required information - textual or visual has become a very tricky task. According to Datta et al. (2008), although there have been advancements and success in textual information retrieval, very little has been achieved in the image retrieval domain. The image based query engines are still immature and rare and need to be further developed.

Currently, most of the research in image retrieval is merely single image query based methods. Content based image retrieval (CBIR) techniques uses a query model specified by the user through an image example or feature. Nonetheless, since the information required by the user usually cannot be represented through a single image, CBIRs' capability is limited as they do not provide any scheme for formulating information that represents multiple examples. However, there are CBIR approaches such as relevance feedback which utilizes multiple examples as a query, but they may require several user feedbacks in order to obtain a refined query, which is inconvenient. Consider a scenario in law enforcement agency, where the investigator needs to search for multiple fingerprints in a dataset. It becomes a highly inconvenient and time - consuming process when the investigator has to look for the fingerprints match one at a time, since only one image (i.e. fingerprint) can be queried at a given time.

On comparison with similar work done by (Basak, Bhattacharya & Chaudhury, 2006), which focuses on retrieval of those images whose features represent the combination of multiple example query content, our method exploits the discriminative different in the multi-example query for retrieval of images.

Active learning technique plays an important role in classification as it actively selects

distinctive information. The advantage of active over passive learning is that it performs selective sampling ensuring the learning process is robust against noise and data scarcity problems (Ling & Du, 2008). Due to active learning's adaptive and evolving characteristics it is potentially useful for targeted learning tasks and works well particularly on large and nonlinear datasets. Currently, active learning has been successfully implemented in fields such as internet security (Long et al., 2008; Almgren & Jonsson, 2004), bioinformatics (Danziger et al., 2007) and text classification (Tong & Koller, 2002).

In the field of image retrieval and recognition, the concept of active learning has only been explored recently. Most of the existing image recognition and retrieval systems utilize single-image querying approach. Moreover, a majority of these classical approaches make use of statistical classifiers for classifying the images. The benefit of these classifiers is that they can be highly accurate when trained on a large database. However, they are computationally expensive and cannot adapt to new data encountered in testing. Also, due to the large number of variables that represent the features of the images, there is no guarantee that the significant features necessary for image recognition have been taken into account.

Taking these issues into consideration, we have proposed an active learning approach based multi-example image retrieval (MeIR) that retrieves the images by exploiting the discriminative features present in the multi-example query. MeIR is adaptive, learns incrementally, has better usability, selects important discriminative features through selective sampling (which consists of Nonparametric Discriminant Analysis (NDA), Incremental Nonparametric Discriminant Analysis (IncNDA), and correlation based criterion).

5.1.1 Related Researches and Motivation

Amongst the many image retrieval techniques, *keywords* - based approaches are most commonly used, where the user inputs query in the form of word or sentence (Kherfi et al., 2004). Search engines such as Google and Yahoo are based on these approaches. Even though popular, this technique has its limitations since keywords based image retrieval can be highly subjective depending on the labels annotated to them. The next is *Image or Graphics* based query where the user provides a query image or a

computer generated picture/graphics for retrieval of similar or same image. *Content based Image Retrieval* (CBIR) makes use of ranking system where it ranks images based on their feature similarity (Huang et al., 2008).

And finally, there is the *relevance feedback* based - image retrieval where the user is asked for a feedback multiple times in order to refine the query concept. Once the image features are extracted, it becomes a challenging task to index and match the images from the image query to the dataset. Various approaches have been used to solve this problem. As stated by Datta et al. (2008), some of the common approaches are feature-based matching, object-silhouette-based matching, structural (hierarchically) feature matching, salient feature matching, and learning-based approaches for similarity matching (Wu et al., 2000), (Weber et al., 2000). Recently, image retrieval using relevance feedback (RF) (i.e. from the user) has been used where the query concept is adaptively redefined iteratively for acquiring the image sought by the user (Huang et al., 2008; Rui, Huang, Ortega & Mehrotra, 1998; Fang, Geman & Boujemaa, 2005; Jaimes, Omura, Nagamine & Hirata, 2004).

The image retrieval system, as the name suggests, corresponds to the method or criterion by which the system retrieves the images. The two most commonly used image retrieval methods utilize the k-nearest neighbour approach, which retrieves the k nearest images based on distance. Other methods use a threshold, say ϵ , where the distance of the images to be retrieved should be less than the given value of ϵ . The drawback of the first technique is the selection of k, and moreover the k-retrieved images may not be close to the query. As for the second technique which uses threshold ϵ , it may return large number of irrelevant images or no images at all.

Image retrieval has a wide range of application in many domains. Due to the abundance of available images it has now become a necessity to have image retrieval system that can efficiently retrieve the expected image. Some of the applications of image retrieval are listed below.

These applications include identifying intellectual property (Foo, Zobel, Sinha & Tahaghoghi, 2007; James, Chang, Wang, Li & Wiederhold, 1998), filtering mature content (Forsyth et al., 1996; Fleck, Forsyth & Bregler, 1996), fingerprint recognition (Ratha, Karu, Chen & Jain, 1996), face recognition (K. Martinez, Cupitt, Saunders & Pillay, 2002), DNA matching and shoe sole impressions (Geradts, 2002) etc.

5.2 Single example as an Image Query

In a conventional image retrieval system that uses single image as a query, the images are retrieved using one of the methods mentioned in the above section. For example, let O_j be the number of images to be retrieved where $j = 1, 2, 3, \dots, n$. Y_t be the image features obtained from the image dataset where $t = 1, 2, 3, \dots, k$. Ω the eigen matrix of Y_t and let Q be the single selected query image by the user. Therefore, the O_j retrieved images for a single query image Q from Y_t can be represented as:

$$O_j = \sum_{j=1}^n \left(\min \left(\sum_t |\Omega(Q) - \Omega(Y_t)| \right)^{1/2} \right). \quad (5.1)$$

As seen in the above Eq.5.1, O_j is obtained based on the k-nearest neighbor (using Euclidean distance as a distance metric) from the single query image.

The problem with this classical single query image retrieval system is that the k-retrieved images may not be close to the query (see table A.1, in Appendix A).

Algorithm 1: Image Retrieval using Single sample as an Image Query

Input:

n – number of images to be retrieved.

Q – Single sample selected by User for querying.

Y_t - image features where $t = 1, 2, 3, \dots, k$.

Step 1: Get Ω NDA of Y_t .

Step 2: Apply Ω on Q .

Step 3: Obtain O_j by getting n Nearest Neighbor of Q from Y_t .

Output:

O_j images retrieved for Q

We have addressed this problem with the proposed method explained in the next section.

5.3 Multiple example as an Image Query

Feature extraction and dimensionality reduction (subspace) techniques such as Principal Component Analysis (PCA) and discriminative analysis are commonly used for

face recognition and retrieval. These techniques are used for extracting informative feature vectors that span a subspace of the images.

Moreover, image datasets having affine distortion/transformation (as in dataset 3 having wide baseline stereo object images), the conventional method (which uses Euclidean distance as in Eq.5.1) for obtaining similar images is suboptimal, since it is difficult to establish correspondence between stereo images by comparing regions of a fixed (Euclidean) shape, as their shape is not preserved under affine transformation. Therefore, we have used Pearson's product-moment correlation coefficient (PMCC) as a similarity metric (see table A.1 in Appendix A).

For obtaining *Pearson's product-moment correlation coefficient* (PMCC), we divide the sample covariance between two vectors/variables by the product of their sample standard deviation (Rodgers & Nicewander, 1988; Stigler, 1989), and can be calculated as:

$$\rho_{(x,y)} = \frac{\sum_m \sum_n (x_{mn} - \bar{x})(y_{mn} - \bar{y})}{\sqrt{(\sum_m \sum_n (x_{mn} - \bar{x})^2)(\sum_m \sum_n (y_{mn} - \bar{y})^2)}}, \quad (5.2)$$

where \bar{x} , and \bar{y} denote the mean or average of the x,y elements respectively. The query images $\Omega(Q_i)$ are updated with newly obtained discriminant information from image dataset features $\Omega(Y_t)$. The updated query $\Omega'(Q_i)$ can be represented as:

$$\Omega'(Q_i) = \Omega(Q_i) + \sum_i \sum_t (\rho_{max}(\Omega(Q_i), \Omega(Y_t))), \quad (5.3)$$

where ρ_{max} denotes the maximally correlating n discriminant vectors found in $\Omega(Y_t)$ for query $\Omega(Q_i)$. After obtaining $\Omega'(Q_i)$, an iterative refinement is carried out for N times. The value of N depends on the image dataset, thus, the retrieved images after iterative refinement can be denoted as:

$$O_j = \sum_N \Omega'(Q_i). \quad (5.4)$$

The following is the summary of the algorithm used in our proposed method.

Algorithm 2: Multi-example Image Retrieval on Active Mode Incremental NDA Learning

Input:

n – number of images to be retrieved.

Q_i – Multiple Image Query selected by User where $i = 1, 2, 3, \dots, N$.

Y_t - image features where $t = 1, 2, 3, \dots, N$.

Step 1: Get Ω NDA of Q_i .

Step 2: Apply Ω on Y_t .

Step 3: Get n maximum correlating images of Q_i from Y_t .

Step 4: Incrementally update eigenspace Ω' from selected n .

Step 5: Apply the updated eigenspace Ω' on Q_i .

Step 6: Iterative refinement from Step 2 for updated $\Omega'(Q_i)$.

Step 7: Iterate i times from Step 3 to obtain retrieved images O_j for Q_i .

Output:

O_j images retrieved for Q_i

5.4 Experiments and Discussion

5.4.1 Experimental Setup

Image representation and feature selection is an important step in our data pre-processing. We have resized and converted all the face images into a 60x60 pixels greyscale images. The obtained pixel values (2 dimensional matrix of 60 x 60) of each image were vectorized into a 1 dimensional row matrix (1 x 3600).

Descriptive features extracted from PCA outperform NDA in some classification tasks and according to Vo, Vo, Challa and Moran (2009), it is less sensitive to different training data sets. Therefore, by combining descriptive features of PCA and discriminant features of NDA, a better performance for dimensionality reduction and feature extraction is achieved. Hence, PCA was applied beforehand, and then the data was normalized. Nonparametric Discriminant Analysis (NDA) was performed, only on queried data in order to obtain the initial NDA eigenspace representation of the face images. For all the experiments, the value of k in NDA and IncNDA has been set to 3.

In this experiment, as a performance metrics we used error rate (calculated based on the number of irrelevant images retrieved, relative to the total number of relevant

images present in the dataset) and recall / percentage accuracy (calculated based on the number of relevant images retrieved, relative to the total number of relevant images present in the dataset) calculated as:

$$ErrorRate = \frac{\text{Number of irrelevant images retrieved}}{\text{Total number of relevant images in the database}}, \quad (5.5)$$

$$Recall/Accuracy = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in the database}}. \quad (5.6)$$

In our experiment we have carried out performance evaluation on three datasets. The images of the individual included in both face datasets (i.e. 1 and 2) have different features/characteristics such as gender, race, age, glasses, beards. While dataset 3 is an object category dataset, consisting of stereo object images.

In face image recognition and retrieval, performance is dependent on several varying factors such as facial illumination, expression, pose etc (Young & Rhee, 2008). We selected datasets 1 and 2, since they consist of facial feature variations due to the change of facial expressions. Dataset 2, apart from facial feature variations, also contains *illumination problem*, which according to Adini, Moses and Ullman (1997); Roy-Chowdhury and Xu (2007), remains a persistent problem in face recognition.

We selected dataset 3 since it consists of wide-baseline stereo problem (where the problem lies in establishing correspondences between a pair of images taken from different viewpoints) (Matas, Chum, Urban & Pajdla, 2004). In the following experiments, we used PMCC for both the Single-example and Multi-example methods, for an unbiased comparison.

Dataset 1: Face Image Dataset This face image dataset consists of 20 face images for each of the 100 individuals making a total of 2000 images. These images were taken from <http://www.essex.ac.uk> and Knowledge Engineering + Discovery Research Institute (KEDRI) <http://www.kedri.info>. Fig.5.1 shows the sample face images belonging to dataset 1. It can be seen that the level of difficulty for recognizing the images is moderate, since the available images for each individuals where taken under the same lighting conditions, however they have noticeable variations in facial expressions and pose.

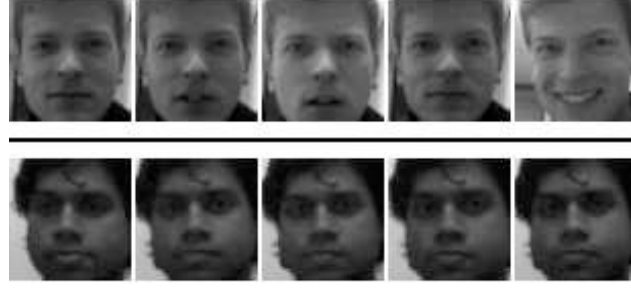


Figure 5.1: Sample images from dataset 1: this dataset has face images with different facial expressions but have uniform illumination

Dataset 2: Light AR Face Recognition Dataset The AR face recognition dataset has been taken from [http : //rvl1.ecn.purdue.edu/v1/ARdatabase/](http://rvl1.ecn.purdue.edu/v1/ARdatabase/) (A. Martinez & Benavente, 1998). This dataset consists of a total 845 images. There are 5 images for each of 169 individuals making it relatively sparse, given the number of images present for each individual. Fig. 5.2 shows the sample face images belonging to the dataset 2. It can be seen that the level of difficulty for recognizing the images when compared to dataset 1 is hard, since each of the image available for every individual have different facial expressions and lighting conditions (varying illumination).



Figure 5.2: Sample images from dataset 2: this dataset has face images with different facial expressions and lighting conditions

Dataset 3: Amsterdam Library of Object Images (ALOI) Dataset The AOLI dataset has been taken from [http : //staff.science.uva.nl/ aloi/](http://staff.science.uva.nl/aloi/) (Geusebroek et al., 2005). The dataset consists of 750 wide baseline stereo object images. However, we have restricted ourselves to 21 randomly selected object categories, for

computational reasons. Wide baseline stereo images are images of a scene or object taken from two arbitrary viewpoints. Given two images of a scene or object taken from arbitrary viewpoints, establishing a dependable association amongst them is a fundamental problem in many computer vision tasks (Matas et al., 2004).



Figure 5.3: *Sample AOLI wide baseline stereo object image. The combination of left-center and center-right images yields two pairs of 15 degree baseline stereo, and the left-right pair combination yields a 30 degree baseline stereo pair. (Geusebroek et al., 2005)*

We have used this dataset to show the robustness of our proposed method against affine distortion/transformation. Unlike traditional approaches; used for stereo image matching, we do not consider any prior knowledge about the relative camera positions, orientations, rotational or affine invariant features for computation.



Figure 5.4: *Sample images from dataset 3: AOLI wide baseline stereo object images dataset*

5.4.2 Case Study 1: Face Image Retrieval

Effects of varying number of query images

In this subsection, the retrieval performance of MeIR is examined for different number of multiple examples representing different features. The performance is demonstrated in Fig. 5.5(a)-(b). In Fig. 5.5(a), we have provided query images of two individuals each having unique features such as glasses or beard. We also note that the images belonging to another individuals have being retrieved since they share similar discriminative features with the multi-example query, most visually noticeable feature being the beard. Similarly, as demonstrated in Fig. 5.5(b), one more query example of individual having both beard and glasses is added for MeIR learning showing that images have more distinctive/discriminant features result in higher retrieval accuracy.

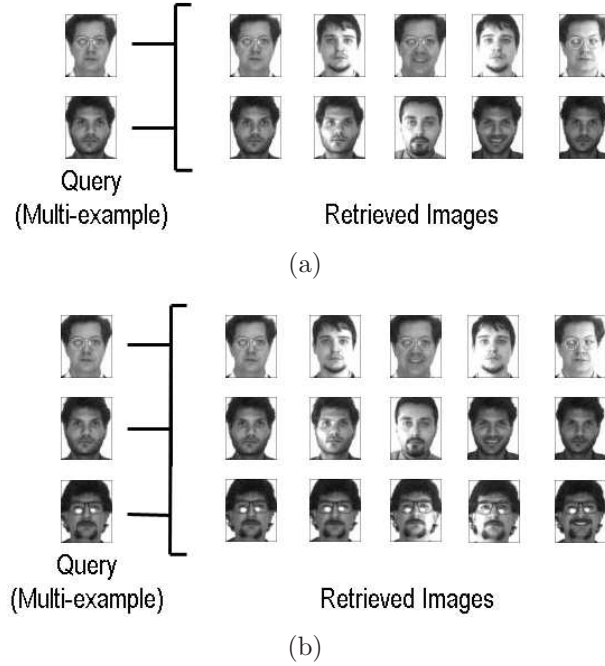


Figure 5.5: Dataset 2: Comparison on varying number of query images. (a) shows the retrieved images from two query images with one individual having glasses and the other having beard. In the retrieved images, we find images of same individuals along with other individuals having similar discriminative features such as beard or glasses. In (b), we add one more query with individual having both beard and sunglasses.

Table 5.1: Performance evaluation using different number of query examples in multi-example method for (face image) dataset 1 and 2. Since the number of individuals in dataset 1 is 100, N/A denotes no data(individual image) available.

No. of Query Examples	Classification Accuracy %	
	Dataset 1	Dataset 2
10	99.89% \pm 0.21	46.00% \pm 8.24
20	99.00% \pm 1.04	44.40% \pm 5.55
30	95.84% \pm 0.80	42.79% \pm 4.25
40	95.68% \pm 0.96	44.80% \pm 4.97
50	95.61% \pm 1.15	46.00% \pm 5.13
60	95.56% \pm 1.15	45.79% \pm 5.58
70	95.46% \pm 0.94	45.14% \pm 5.51
80	95.84% \pm 0.81	44.75% \pm 5.44
90	95.58% \pm 0.77	44.66% \pm 5.56
100	96.70% \pm 0.66	45.40% \pm 5.54
110	N/A	45.71% \pm 5.23
120	N/A	45.89% \pm 5.29
130	N/A	45.57% \pm 5.05
140	N/A	45.31% \pm 4.83
150	N/A	46.02% \pm 4.65
160	N/A	46.41% \pm 4.50
169	N/A	46.86% \pm 4.74

Table 5.1 shows the percentage accuracy along with confidence interval for face dataset 1 and 2. In case of dataset 1, where 20 images are present for each individual, the percentage accuracy was calculated using 20 - fold cross validation. Similarly, for dataset 2 that contains five images per individual, 5 fold cross validation was used. It can be observed that multi-example method performs gracefully under different number of query examples, without significant impact on performance.

Single-example method versus Multi-example method

Fig.5.6 illustrates the incorrectly retrieved images encapsulated in frames, where six images of individuals from dataset 2 were chosen randomly for testing the percentage of error rate in image retrieval for each individual. For the single-example method,
















































































Single-example Method		Multi-example Method	
Query	Retrieved Images	Query	Retrieved Images
	     		    
	     		    
	     		    
	     		    
	     		    
	     		     

Figure 5.6: Images retrieved by single-example method and multi-example method from dataset 2. The images encapsulated in frames are incorrectly retrieved face images.

the six images were provided one at a time, whereas for the proposed multi-example methods all the six images were queried in one pass. The overall error rate for the single-example was 16.4%, surprisingly for the proposed multi-example method it was 6.42%, which is 9.98% less than the single-example method. This experiment demonstrates the ability of our proposed method to process all the examples at once in one pass based on the discriminative information among the queried multi-example face images.

Iterative refinement is the term we have used to describe the number of iterations that are required in order to obtain the optimal solution. Fig.5.7 shows that initially when the number of iterations is 5, the error rate is 15%. Progressively, as the number of iterations increases, the error rate gradually starts dropping since the incremental NDA learns and updates the eigen model from the newly obtained relevant images with each iteration. After 20 iterations it can be seen that error rate stabilizes at 6.4%. The number of iterations required for a particular dataset depends on the number of instances per class. For datasets 2 and 3, the number of instances available per class (or number of images per person/object) is five and three respectively. The number of iterative refinement required for both dataset 2 and dataset 3 is 7.

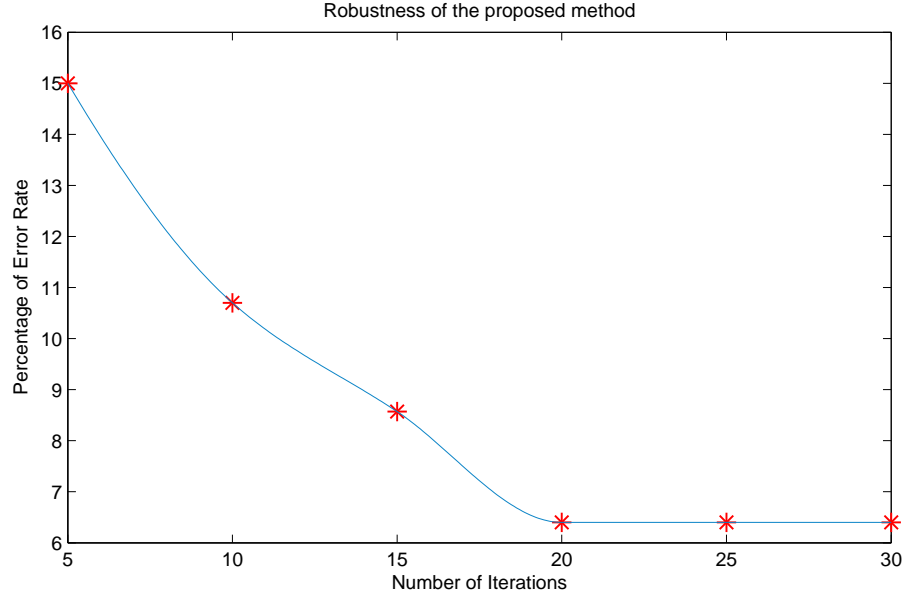


Figure 5.7: Performance evaluation (in terms of error rate versus number of iterations) of the proposed multi-image query based image retrieval on dataset 1. It can be seen that after 20 iterations the error rate stabilizes at 6.4%

Table 5.2 shows the overall performance evaluation of single-example method based face image retrieval against the proposed multi-example method. It can be seen that aIncNDA performs optimally on dataset 1. Whereas, for dataset two (due to the illumination problem), the multi-example method's performance is similar to that of single-example method.

Table 5.2: Overall percentage accuracy of single-example and multi-example method for face datasets.

Method	Classification Accuracy %	
	Dataset 1	Dataset 2
Multi-example	96.70 ± 0.66	46.86 ± 4.74
Single-example	56.91 ± 23.34	46.43 ± 5.26

From Fig.5.9 and Table 5.2, it can be observed that for dataset 1, the (multi-example) proposed method performs very well with an accuracy of $96.70\% \pm 0.66$, whereas the single-example method, which uses passive nonparametric discriminant analysis (NDA), has a retrieval accuracy of $56.91\% \pm 23.34$. However, for dataset 2 the single-example method and multi-example method perform almost equally.

5.4.3 Case Study 2: Object Image Retrieval

As seen in table 5.3, for dataset 3 our proposed method has an accuracy of $73.53\% \pm 3.66$, which is significantly better than Single-example methods, which has an accuracy of $48.67\% \pm 22.26$.

Table 5.3: Overall percentage accuracy of single-example method and multi-example method for dataset 3.

Dataset 3: ALOI	
Method	Classification Accuracy (%)
Multi-example	73.53 ± 3.66
Single-example	48.67 ± 22.26

Fig. 5.9 graphically illustrates the overall performance evaluation (in terms of percentage accuracy) of the proposed multi-example method based image retrieval (MeIR) versus single-example method, for all three image datasets. Through this wide baseline stereo object image dataset, we have thus demonstrated the robustness of MeIR against affine distortion/transformation.

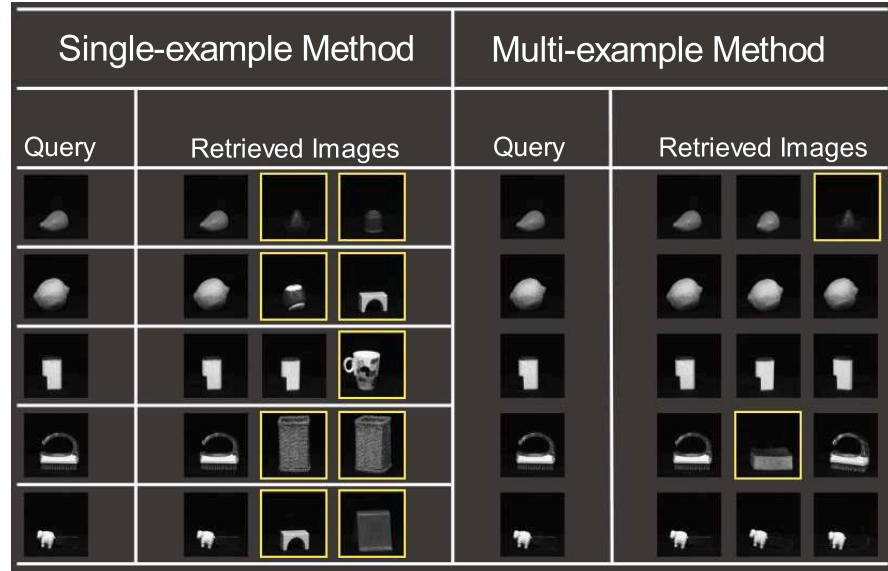


Figure 5.8: Images retrieved by single-example method and multi-example method from dataset 3. The images encapsulated in frames are incorrectly retrieved object images.

Unlike conventional approaches (used for stereo image matching), we do not consider

any prior knowledge about the relative camera positions, orientations, rotational or affine invariant features for computation. It is difficult to establish correspondence between stereo images by comparing regions of a fixed (Euclidean) shape, since their shape is not preserved under affine transformation, even then MeIR's performance is near optimal due to its adaptive and incremental learning nature.

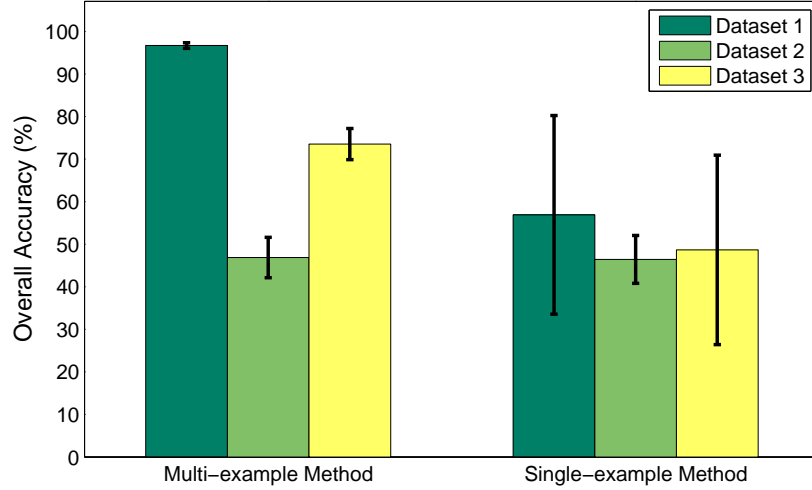


Figure 5.9: Overall performance evaluation of the proposed multi-example method based image retrieval.

5.4.4 Discussion

In general, the illumination problem is quite difficult and has received consistent attention in the image processing domain. From Fig.5.11, it can be observed how the projection vectors undergo change due to illumination. Given the sparsity of instances for each class in dataset 2, and each image being different in terms of luminosity. The proposed multi-example method's performance was similar to the single-example method.

This may be due to the fact that for face recognition we have used only discriminative features; therefore multi-example method works optimally where sufficient number of clear discriminative features are present in the image dataset (as in Dataset 1). It would be interesting to carry out an experiment where both geometric and photometric features are considered (for active and discriminative incremental learning) with

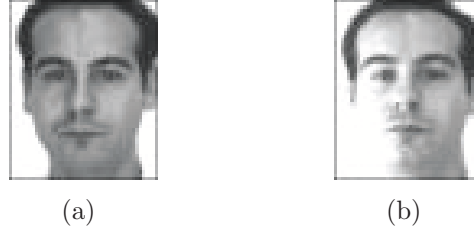


Figure 5.10: Dataset 2: Illumination Problem. (a) Image under uniform lighting/luminosity. (b) Image under lighting/luminosity focus from left side.

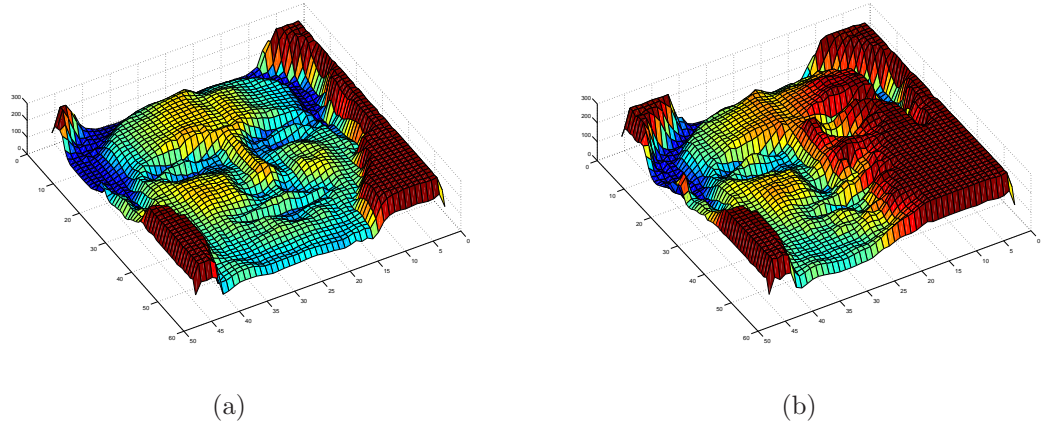


Figure 5.11: Dataset 2: Three-dimensional surface plot of images from Fig.5.10. The above figure shows how projection vectors are changed due to illumination/lighting conditions.

reasonable number of face images per individual. Also, pre-processing the images for illumination restoration or light intensity (luminosity) equalization would prove to be beneficial for obtaining better recognition and retrieval.

For experimentation, we have selected three different datasets, each having different problems such as facial expression, illumination, pose, affine distortion and rotation, and works significantly well on all problems; apart from the image dataset having illumination problem (as in dataset 2). From the above results it can be observed that the multi-example active learning system dynamically and incrementally learns from the newly obtained images thereby gradually reducing its error rate by means of iterative refinement. Furthermore, maximum of only one image is taken from selected classes (obtained from query) for training (building of initial eigenspace). This shows that our proposed method requires the least number of images per class when compared to traditional methods.

However, for efficient and successful image retrieval, the selection of features that represents the image similarity and dissimilarity is very important since the success of relevant image retrieval is highly dependent on these factors.

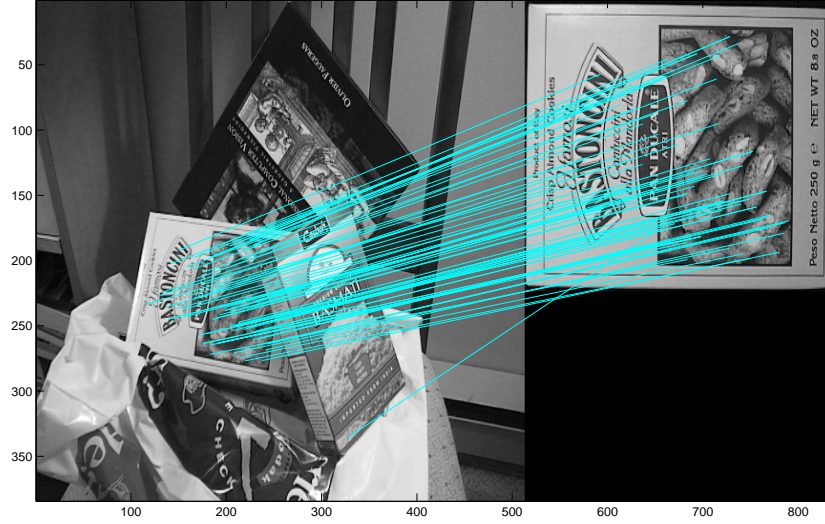


Figure 5.12: Example of image recognition using Affine Scale-Invariant Feature Transform (ASIFT) features. Amongst the key points found in both images using Harris-Affine method, 80 matches were found. The matching affine invariant key points between the two images are shown in the above figure.

Therefore, in future work new selective sampling techniques based on Affine Scale-Invariant Feature Transform (ASIFT) features will be exploited. ASIFT features (Morel & Yu, 2009; Vedaldi, 2007) are fully invariant to parameters such as zoom, rotation, translation, and the angles defining the camera axis orientation as shown in Fig.5.12. Moreover, the method permits to reliably identify features that have undergone very large affine distortions or transition tilt using Harris-Affine and Hessian-Affine techniques. However, on experimentation, it was found that ASIFT features only allows recognition of same images under different rotation and affine distortions, but not similar images making it unsuitable for datasets such as Caltech - 101 which has a large amount of inter-class variability.

As future work, our proposed method will be incorporated with ASIFT features (vectors) and bag-of-features representation models (Li & Perona, 2005), and test it on objects based image dataset such as Caltech-101/256.

5.5 Summary

In this chapter we applied our active learning methods (aIncNDA) for image recognition and retrieval and have further discussed the advantages and limitations which are inherent in it. With the next chapter, we will be concluding this thesis stating the limitations and future works.

Chapter 6

Conclusion and Future Works

*When you have eliminated the impossible,
whatever remains,
however improbable, must be the truth.*

*- Sherlock Holmes,
The Sign of Four.*

6.1 Conclusion

A method based on passive learning proves to be inadequate in real world application. Active learning could potentially empower other techniques/methods with flexibility and efficiency. This thesis introduces a novel active mode incremental nonparametric discriminant analysis (aIncNDA) learning method, in which the passive incremental NDA is extended with data selective sampling, and performs active online discrimination analysis.

Given an incoming instance, the aIncNDA computes a discrimination residue ratio between within-class and between-class, in which the residue is calculated using the k^{th} regional nearest neighbor to class mean vector. The proposed aIncNDA is capable of estimating the discriminant contribution for every newly presented instance, because the discrimination residue rate imitates the fundamental NDA criterion for a maximum separation between classes and minimum separation within classes. In our experiments, we described how the discriminative instances can be

significantly selected based on discrimination residue with, at most, minor sacrifices in learning rate and classification accuracy.

The experimental results show that the proposed aIncNDA performs gracefully under different level of redundancy, and the aIncNDA learning system is capable of learning with fewer instances, but has more often an improved discrimination performance, than a passive incremental NDA.

6.1.1 Contributions

The proposed active mode incremental nonparametric discriminant analysis(aIncNDA) methods contributions can be summarized as:

It is an unsupervised, active and incremental learning method that automatically selects only those instances that are beneficial for the targeted learning task. This unsupervised selective sampling reduces the computational cost of training the classifier on all historical data by selecting fewer (discriminative) instances.

The incremental learning nature of the proposed methods reduces the risk of concept drift. Moreover, incremental learning reduces the computational cost by updating the eigen model without having to compute from historical data again. aIncNDA also works well on large, streaming, and constantly changing data due to its adaptive and incremental learning nature.

6.1.2 Limitations

The data processing in aIncNDA is not one-pass, and although aIncNDA facilitates online (adaptive) learning, it is unable to carry out the task in real - time. In order to facilitate real time adaptive learning, a more efficient criterion needs to be introduced.

The classifier used in the aIncNDA needs to be retrained each time new data is introduced, and the parameters of the classifier are not optimized. To solve this problem, an incremental learning classifier needs to be introduced; along with decremental learning for optimal performance.

For incremental and batch NDA, how many nearest neighbor should be selected for a particular dataset or new incoming instances for optimal performance.

The used SVM and k-NN classification methods, integrated with aIncNDA are limited in their capacity to incrementally learn and accurately classify new data.

6.2 Future Directions

Neural networks based methods have recently gained attention in many domains in the context of pattern recognition problems. They have been used for classification, prediction, image processing, anomaly detection etc. Previously, these type of problems were generally solved by (classical) statistical approaches. However, as discussed in the literature review, the parametric statistical approaches suffer from assumptions that the underlying probability distribution for the given data(sets) is linear. And Free and Schumann (1997) states that nonparametric statistical methods are unsuitable for sparse datasets.

There are several instances where neural network based methods have performed better than pure statistical approaches. For example, Free and Schumann (1997) has compared neural networks with several statistical classification methods such as Linear Discriminant Analysis, Quadratic Discriminant Analysis, Discriminant Analysis with nonparametric density estimators and k-NN classification algorithms on well log data. In the experiment, the neural network approach was found to be slightly more advantageous than the statistical approach for classification task. The author states that since both the methods (neural networks and statistical approaches) aim to reach the Bayes error based on unknown probability distribution of the data, the parametric statistical methods can benefit if there is reasonable assumption about the probability distribution. Otherwise, neural networks approach will be useful since they directly seek for discriminating surfaces without prior knowledge of the datasets probability model.

Sheel, Vrooman, Renner and Dawsey (2001) in their study have compared neural networks approach with Fishers Linear Discriminant Analysis (FLDA). The results show that neural networks outperform classical discriminant analysis in prediction

task. However, in the study conducted by Wilson and Hardgrave (1995), the discriminant analysis approach performed better. The justification provided by the authors is that discriminant analysis approach is favourable where the groups to be discriminated are linearly separable.

Similarly, Cooper (1999), Ripley (1994) and Legitimus and Schwab (1991) have also compared neural network with several multivariate statistical techniques. A comparison of the results with those obtained from multivariate statistical procedures applied to the same data set suggests that neural networks are worthy of consideration as a potentially valuable complementary tool along with statistical techniques in the machine learning domain.

In further works, the current active learning model will be extended by incorporating techniques such as incremental classifiers (probabilistic spiking neural networks) and feature selection. Also more efficient criterion will be explored and exploited for selective sampling technique.

Automated optimization, using methods involving spiking neural networks, for classifier's parameters and determining the number of nearest neighbor to be selected for a particular dataset; or new incoming instances will be a major focus for future works.

Inspired by Kasabov (2009a), active learning using evolving probabilistic spiking neural networks utilizing quantum inspired evolutionary algorithm shall be explored.

Active learning models inspiration comes from the cognitive learning ability of the human brain, whereas spiking neural network models inspiration comes from the spiking processes in biological neurons. Thus, integrating these two models together would enable a more realistic mimicry of the human's innate learning ability. Therefore, based on this idea, I would like to study for a PhD degree the following:

- pSNN as incrementally evolving classifiers for active learning (instead of SVM and kNN)
- Quantum superposition as feature representation and feature selection for image classification and image associative memories.

6.2.1 Incremental evolving probabilistic spiking neural networks (pSNN) for active learning.

Compared to traditional Neural Networks, SNN requires less number of neurons and works well with spatio-temporal data (Maass & Bishop, 1999; Gerstner & Kistler, 2002). I will develop an active learning model based on probabilistic spiking neural networks (pSNN). Amongst the neural network models, spiking neural networks (SNN) mechanism are more realistic in terms of spiking processes to the biological neurons (Maass & Bishop, 1999). According to Kasabov (2009b), since “the spiking processes in biological neurons are stochastic by nature it would be appropriate to look for new inspirations to enhance the current SNN models with probabilistic parameters”. Therefore, motivated by Kasabov (2009b, 2009a), developing an active learning model based on evolving probabilistic spiking neural networks (pSNN) should result in efficient learning and classification of new data.

6.2.2 Quantum superposition as feature representation and feature selection for image classification and image associative memories.

Feature representation and feature selection are common preprocessing steps in the machine learning domain. It is a necessity since; it allows us to efficiently counter the ‘curse of dimensionality’ for dataset have high dimensions. It is a challenging task to extract appropriate features, especially when the amount of data that needs to be processed is massive. In these cases, the conventional feature representation and feature selection methods (such as PCA, LDA and NDA) do not perform optimally.

Traditional methods like “look up tables” or “hash-tables” are not optimal compared to neural networks because access to the patterns is slow if the look up table contains too many patterns. The fault tolerance cannot be easily implemented in hash-tables. And for n^2 stored patterns, n^2 times the number of steps are required for comparing two patterns.

Further, Knoblauch (2005, 2004) in their study have shown that SNN improves fault tolerance against noise and allows fast separation of superpositioned patterns by

making use of precise spike timing. Also, Izhikevich (2006) states that using SNN increases the associative memories capacity.

For need of an optimal feature representation and feature selection, inspired by Kasabov (2009a), Schliebs, Defoin-Platel and Kasabov (2009) for my PhD studies, I would like to explore quantum superposition as feature representation and feature selection for image classification and image associative memories.

Associative memories are related to the human ability to retrieve information from applied associated stimuli. Modeling and incorporating this ability using quantum superposition principles would be a crucial addition to the *active learning model based on evolving probabilistic spiking neural networks*, as it would allow an adeptly evolving incremental learning within a limited memory space. A recent manifestation of quantum inspired evolutionary algorithms (QiEA), by Kasabov (2009a); Schliebs et al. (2009), have made use of the quantum principle of superposition. The authors state that a bit (which is the smallest information unit in digital computers), exists in either ‘1’ or ‘0’ states at any given time. Similarly, in quantum physic, a quantum bit (qbit) can exists in ‘1’ or ‘0’ states, but also in a superposition of both states, and can be represented as:

$$|\Psi\rangle = \alpha |0\rangle + \beta |1\rangle, \quad (6.1)$$

where α and β are complex number that are used to define a qbits probable state $|\Psi\rangle$.

References

- Abe, N., Zadrozny, B. & Langford, J. (2006). Outlier detection by active learning. In *Kdd '06: Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 504–509). New York, NY, USA: ACM.
- Adini, Y., Moses, Y. & Ullman, S. (1997). Face recognition: the problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 721–732.
- Almgren, M. & Jonsson, E. (2004, June). Using active learning in intrusion detection. In *Proceedings of the 17th ieee computer security foundations workshop*. (p. 88–98).
- Basak, J., Bhattacharya, K. & Chaudhury, S. (2006, Dec.). Multiple exemplar-based facial image retrieval using independent component analysis. *IEEE Transactions on Image Processing*, 15(12), 3773–3783.
- Cooper, J. C. B. (1999). Artificial neural networks versus multivariate statistics: an application from economics. *Journal of Applied Statistics*, 26(8), 909–921.
- Cormen, T. T., Leiserson, C. E. & Rivest, R. L. (1990). *Introduction to algorithms*. Cambridge, MA, USA: MIT Press.
- Danziger, S. A., Zeng, J., Wang, Y., Brachmann, R. K. & Lathrop, R. H. (2007). Choosing where to look next in a mutation sequence space: Active learning of informative p53 cancer rescue mutants. *Bioinformatics*, 23, i104–i114.
- Dasgupta, S. & Hsu, D. (2008). Hierarchical sampling for active learning. *Proceedings of the 25th international conference on Machine learning*, 12, 208–215.
- Datta, R., Joshi, D., Li, J. & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2), 1–60.
- Denning, D. E. (1987). An intrusion-detection model. *IEEE Trans. Soft. Eng.*, 13(2), 222–232.
- Fang, Y., Geman, D. & Boujemaa, N. (2005). An interactive system for mental face retrieval. In *Mir '05: Proceedings of the 7th acm sigmm international workshop on multimedia information retrieval*. (pp. 193–200). New York, NY, USA: ACM.
- Fleck, M. M., Forsyth, D. A. & Bregler, C. (1996). Finding naked people. In *Eccv '96: Proceedings of the 4th european conference on computer vision-volume ii*.

- (pp. 593–602). London, UK: Springer-Verlag.
- Foo, J. J., Zobel, J., Sinha, R. & Tahaghoghi, S. M. M. (2007). Detection of near-duplicate images for web search. In *Civr '07: Proceedings of the 6th acm international conference on image and video retrieval*. (pp. 557–564). New York, NY, USA: ACM.
- Forsyth, D., Malik, J., Fleck, M., Greenspan, H., Leung, T., Belongie, S. et al. (1996). *Finding pictures of objects in large collections of images* (Tech. Rep.). Berkeley, CA, USA.
- Free, A. S. & Schumann, A. (1997). Neural networks versus statistics: A comparing study of their classification performance on well log data. In *Proceedings of ianng'97, part* (pp. 237–241).
- Freund, Y., Seung, H. S., Shamir, E. & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3), 133–168.
- Geradts, Z. (2002). *Content-based information retrieval from forensic image databases*. Unpublished doctoral dissertation, Utrecht University, The Netherlands.
- Gerstner, W. & Kistler, W. M. (2002). *Spiking neuron models* (1st ed.). Cambridge University Press.
- Geusebroek, J., Burghouts, G. & Smeulders, A. (2005). The amsterdam library of object image. *International Journal of Computer Vision.*, 61, 103–112.
- Greiner, R., Grove, A. J. & Roth, D. (2002). Learning cost-sensitive active classifiers. *Artificial Intelligence.*, 139(2), 137–174.
- Hettich, S. & Bay, S. D. (1999). *The UCI KDD archive*. <http://kdd.ics.uci.edu>. Irvine, CA.
- Hoi, S. C. H., Jin, H. R. & Lyu, M. R. (2006). Large-scale text categorization by batch mode active learning. *Proceedings of the 15th international conference on World Wide Web.*, 633–642.
- Hoi, S. C. H. & Lyu, M. R. (2005). A semi-supervised active learning framework for image retrieval. In *Cvpr '05: Proceedings of the 2005 ieee computer society conference on computer vision and pattern recognition*. (Vol. 2, pp. 302–309). Washington, DC, USA: IEEE Computer Society.
- Huang, T., Dagli, C., Rajaram, S., Chang, E., Mandel, M., Poliner, G. et al. (2008, April). Active learning for interactive multimedia retrieval. *Proceedings of the IEEE*, 96(4), 648–667.
- Izhikevich, E. M. (2006). Polychronization: Computation with spikes. *Neural Com-*

- put.*, 18(2), 245–282.
- Jaimes, A., Omura, K., Nagamine, T. & Hirata, K. (2004). Memory cues for meeting video retrieval. In *Carpe'04: Proceedings of the the 1st acm workshop on continuous archival and retrieval of personal experiences*. (pp. 74–85). New York, NY, USA: ACM.
- James, E. C., Chang, E. Y., Wang, J. Z., Li, C. & Wiederhold, G. (1998). Rime: A replicated image detector for the world-wide web. In *Proceedings of spie symposium of voice, video, and data communications* (pp. 58–67).
- Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Icml '99: Proceedings of the 16th international conference on machine learning*. (pp. 200–209). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Kaelbling, L. P., Littman, M. L. & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237–285.
- Kai, Y., Zhu, S., Xu, W. & Gong, Y. (2008). Non-greedy active learning for text categorization using convex transductive experimental design. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 635–642.
- Kapoor, A. & Greiner, R. (2005). Learning and classifying under hard budgets. In (Vol. 3720, pp. 170–181). Springer Berlin.
- Kasabov, N. (1996). *Foundations of neural networks, fuzzy systems, and knowledge engineering*. Cambridge, MA, USA: MIT Press.
- Kasabov, N. (2009a). Integrative probabilistic spiking neural networks utilizing quantum inspired evolutionary algorithm: A computational framework. *ICO-NIP '08: 15th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Assembly*, 5506, 3–13.
- Kasabov, N. (2009b, September). To spike or not to spike: A probabilistic spiking neuron model. *Neural Networks*.
- Kasabov, N., Middlemiss, M. & Lane, T. (2003). A generic connectionist-based method for on-line feature selection and modelling with a case study of gene expression data analysis. In *Apbc '03: Proceedings of the first asia-pacific bioinformatics conference on bioinformatics*. (pp. 199–202). Darlinghurst, Australia, Australia: Australian Computer Society, Inc.
- Kherfi, M. L., Ziou, D. & Bernardi, A. (2004). Image retrieval from the world

- wide web: Issues, techniques, and systems. *ACM Computing Surveys (CSUR)*, 36(1), 35–67.
- Kim, M., Kim, D. & Lee, S. (2003, Nov.). Face recognition using the embedded hmm with second-order block-specific observations pattern recognition. In (Vol. 36, pp. 2723–2735). Elsevier Ltd.
- Knoblauch, A. (2004). Synchronization and pattern separation in spiking associative memories and visual cortical areas. Available from <http://vts.uni-ulm.de/doc.asp?id=3762>
- Knoblauch, A. (2005). Neural associative memory for brain modeling and information retrieval. *Inf. Process. Lett.*, 95(6), 537–544.
- Kuo, B.-C. & Landgrebe, D. (2004). Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(5), 1096–1105.
- Lee, W., Stolfo, S. & Mok, K. (1999). A data mining framework for building intrusion detection models. *Proceedings of the IEEE Symposium on Security and Privacy.*, 120–132.
- Legitimus, D. & Schwab, L. (1991, Aug). Experimental comparison between neural networks and classical techniques of classification applied to natural underwater transients identification. *Neural Networks for Ocean Engineering*, 113–120.
- Lewis, D. D. & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Sigir '94: Proceedings of the 17th annual international acm sigir conference on research and development in information retrieval* (pp. 3–12). New York, NY, USA: Springer-Verlag New York, Inc.
- Li, F.-F. & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Cvpr '05: Proceedings of the ieee computer society conference on computer vision and pattern recognition*. (Vol. 2, pp. 524–531). Washington, DC, USA: IEEE Computer Society.
- Ling, C. X. & Du, J. (2008). Active learning with direct query construction. In *Kdd '08: Proceeding of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 480–487). New York, NY, USA: ACM.
- Liu, Y. (2004). Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inf. Comput. Sci.*, 44, 1936–1941.
- Long, J., Yin, J.-P., Zhu, E. & Zhao, W.-T. (2008, July). A novel active cost-sensitive learning method for intrusion detection. In (Vol. 2, p. 1099–1104).

- Maass, W. & Bishop, C. M. (Eds.). (1999). *Pulsed neural networks*. Cambridge, MA, USA: MIT Press.
- Martinez, A. & Benavente, R. (1998). The AR face database. *CVC Tech. Report #24*.
- Martinez, K., Cupitt, J., Saunders, D. & Pillay, R. (2002, Jan). Ten years of art imaging research. *Proceedings of the IEEE*, 90(1), 28-41.
- Matas, J., Chum, O., Urban, M. & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10), 761-767. Available from <http://www.sciencedirect.com/science/article/B6V09-4CPM632-1/2/7e4b5f8aa5a4d6df0781ecf74dfff3c1> (British Machine Vision Computing 2002)
- Melville, P. & Mooney, R. J. (2004). Diverse ensembles for active learning. *Proceedings of the 21st international conference on Machine learning*.
- Morel, J. & Yu, G. (2009). Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2), 438-469.
- Nguyen, H. & Smeulders, A. (2004). Active learning using pre-clustering. *Proceedings of the 21st international conference on Machine learning*.
- Pang, S., Dhoble, K., Chen, Y., Kasabov, N., Ban, T. & Kadobayashi, Y. (2009, July). Active mode incremental nonparametric discriminant analysis learning. In *Ims '09: Proceedings of the 8th international conference on information and management sciences*. (pp. 407-412). Kunming, China.
- Pang, S. & Kasabov, N. (2004, July). Inductive vs transductive inference, global vs local models: Svm, tsvm, and svmt for gene expression classification problems. In *Proceedings of the ieee international joint conference on neural networks* (Vol. 2, p. 1197-1202).
- Pang, S., Ozawa, S. & Kasabov, N. (2005). Chunk incremental lda computing on data streams. In *Isnn '05: Advances in neural networks* (pp. 51-56). Springer Berlin / Heidelberg.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 559-572.
- Raducanu, B. & Vitriá, J. (2008). Online nonparametric discriminant analysis for incremental subspace learning and recognition. *Pattern Analysis and Applications*, 11(3-4), 259-268.
- Ratha, N., Karu, K., Chen, S. & Jain, A. (1996, Aug). A real-time matching system

- for large fingerprint databases. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 18(8), 799-813.
- Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3), 409-456. Available from <http://www.jstor.org/stable/2346118>
- Rodgers, J. & Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59-66.
- Roy-Chowdhury, A. K. & Xu, Y. (2007). *Pose and illumination invariant face recognition using video sequences*. Springer Berlin Heidelberg.
- Rubinstein, Y. D. & Hastie, T. (1997). Discriminative vs informative learning. In *Proceedings of the 3rd international conference on knowledge discovery and data mining* (pp. 49-53). AAAI Press.
- Rui, Y., Huang, T., Ortega, M. & Mehrotra, S. (1998, September). Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5), 644-655.
- Schein, A. I. (2005). *Active learning for logistic regression*. Unpublished doctoral dissertation, Philadelphia, PA, USA. (Supervisor-Ungar, Lyle H.)
- Schliebs, S., Defoin-Platel, M. & Kasabov, N. (2009). Integrated feature and parameter optimization for an evolving spiking neural network: Exploring heterogeneous probabilistic models. *Neural Networks*, 22(5-6), 623-632.
- Seung, H. S., Oppen, M. & Sompolinsky, H. (1992). Query by committee. In *Colt '92: Proceedings of the fifth annual workshop on computational learning theory* (pp. 287-294). New York, NY, USA: ACM.
- Sheel, S. J., Vrooman, D., Renner, R. S. & Dawsey, S. K. (2001). A comparison of neural networks and classical discriminant analysis in predicting students' mathematics placement examination scores. In *International conference on computational science (2)* (p. 952-957).
- Siolas, G. & Buc, F. d'Alche. (2000). Support vector machines based on a semantic kernel for text categorization. In *Ijcnnp 2000: Proceedings of the ieee-inns-enns international joint conference on neural networks*. (Vol. 5, p. 205-209 vol.5).
- Stigler, S. M. (1989). Francis galton's account of the invention of correlation. *Statistical Science*, 4(2), 73-79.
- Sugiyama, M. & Nakajima, S. (2009). Pool-based active learning in approximate linear regression. *Machine Learning.*, 75(3), 249-274.

- Symons, C. T., Samatova, N. F., Krishnamurthy, R., Park, B. H., Umar, T., Buttler, D. et al. (2006). Multi-criterion active learning in conditional random fields. *IEEE International Conference on Tools with Artificial Intelligence.*, 323-331.
- Tong, S. (2001). *Active learning: Theory and applications*. Doctoral dissertation, Stanford University, Department of Computer Science.
- Tong, S. & Chang, E. (2001). Support vector machine active learning for image retrieval. In *Multimedia '01: Proceedings of the ninth acm international conference on multimedia* (pp. 107-118). New York, NY, USA: ACM.
- Tong, S. & Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2, 45-66.
- Vapnik, V. N. (1998). *Statistical learning theory*. New York: John Wiley and Sons Inc.
- Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer-Verlag.
- Vedaldi, A. (2007). *An open implementation of the SIFT detector and descriptor* (Tech. Rep. No. 070012). UCLA CSD.
- Vo, N., Vo, D., Challa, S. & Moran, B. (2009). Parametric subspace analysis for dimensionality reduction and classification. In *Cidm '09: Ieee symposium on computational intelligence and data mining* (pp. 363-366).
- Wang, X. & Paliwal, K. (2002, Nov.). Discriminative learning and informative learning in pattern recognition. In *Iconip '02: Proceedings of the 9th international conference on neural information processing*. (Vol. 2, pp. 862-865).
- Warmuth, M. K., Liao, J., Ratsch, G., Mathieson, M., Putta, S. & Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.*, 43(2), 667-673.
- Weber, M., Welling, M. & Perona, P. (2000). Unsupervised learning of models for recognition. In *Eccv '00: Proceedings of the 6th european conference on computer vision-part i*. (pp. 18-32). London, UK: Springer-Verlag.
- Wilson, R. L. & Hardgrave, B. C. (1995). Predicting graduate student success in an mba program: Regression versus classification educational and psychological measurement. *Educational and Psychological Measurement*, 55(2), 186-195.
- Wu, Y., Tian, Q. & Huang, T. (2000). Discriminant-em algorithm with application to image retrieval. *Proceedings of the IEEE Conference on Computer Vision*

- and Pattern Recognition.*, 1, 222–227.
- Young, N. M. & Rhee, P. K. (2008). A novel efficient face recognition using two level evolution classifier. *International Conference on Convergence Information Technology.*, 2, 812-815.
- Yu, K., Bi, J. & Tresp, V. (2006). Active learning via transductive experimental design. In *Icml '06: Proceedings of the 23rd international conference on machine learning* (pp. 1081–1088). New York, NY, USA: ACM.
- Zhang, W. V., He, X., Rey, B. & Jones, R. (2007). Query rewriting using active learning for sponsored search. In *Sigir '07: Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval* (pp. 853–854). New York, NY, USA: ACM.

Appendices

Appendix A

Performance evaluation figures

aIncNDA versus IncNDA performance evaluation figures for benchmark dataset: Some of the dataset graphs from table 4.4 are illustrated below. The graph shows the stages of learning process at every chunk after each update of the initial NDA eigenspace. The graph falls some times, because the percentage of total data being calculated is relative to the number of classes obtained at that particular given stage.

The updated data is provided in a total of 10 stages/chunks. In each stage/chunk 10 percent of the total data is present (randomly) such that the samples are not recurring, therefore each update has unique non-overlapped training samples.

Figure Naming Convention:

$$<DatasetIndex>(<DatasetName>)<NearestNeighborSelected>$$

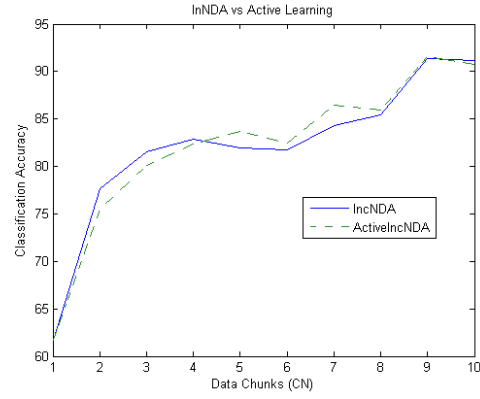


Figure A.1: *D1 (Wisconsin) 2NN*

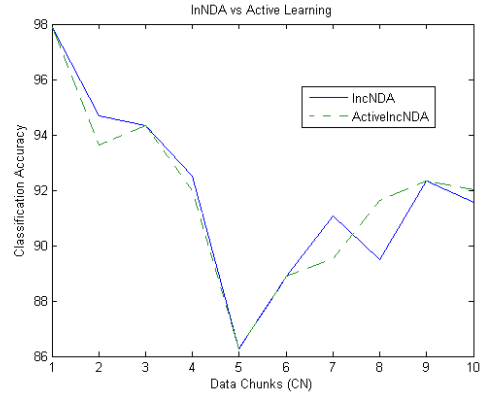


Figure A.2: *D1 (Wisconsin) 5NN*

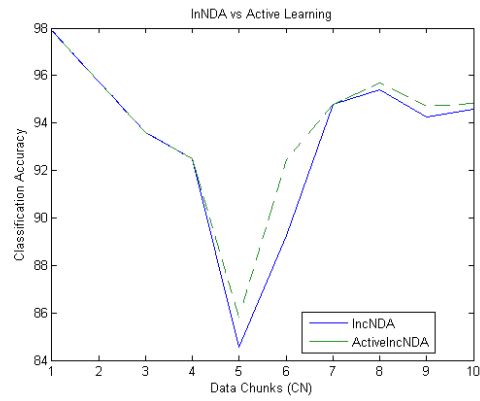


Figure A.3: *D1 (Wisconsin) 7NN*

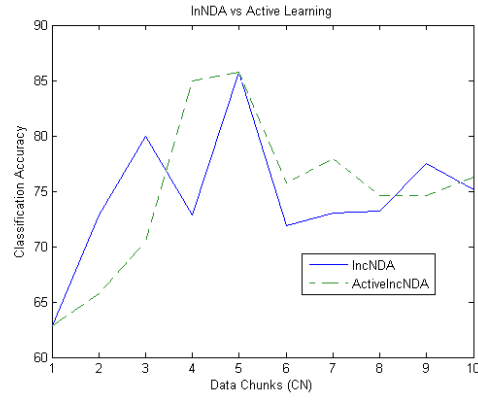


Figure A.4: *D2 (Ionosphere) 2NN*

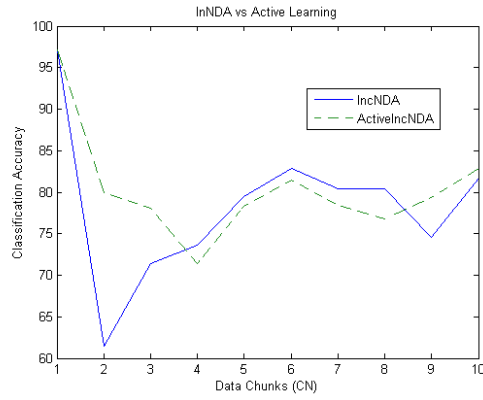


Figure A.5: *D2 (Ionosphere) 5NN*

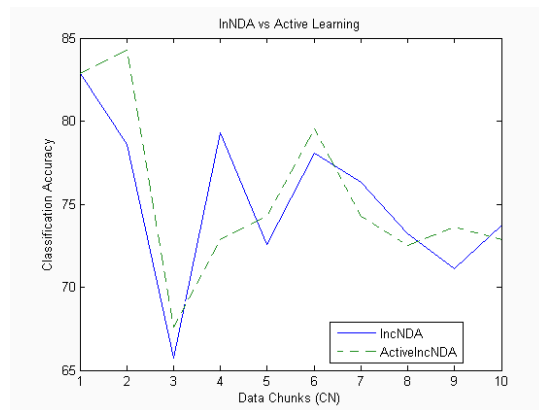


Figure A.6: *D2 (Ionosphere) 7NN*

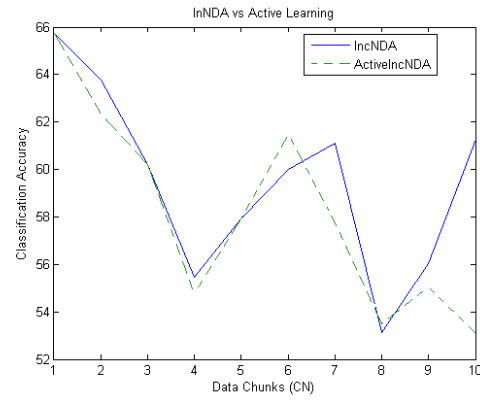


Figure A.7: *D3 (Liver Disorder) 2NN*

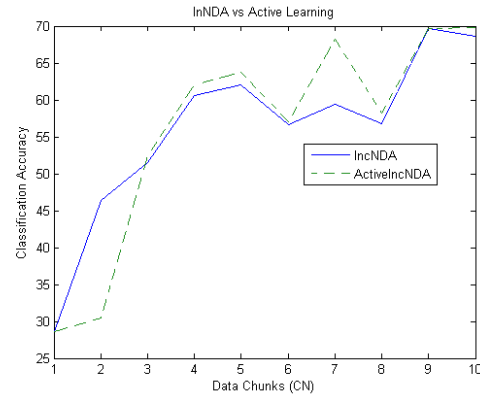


Figure A.8: *D3 (Liver Disorder) 5NN*

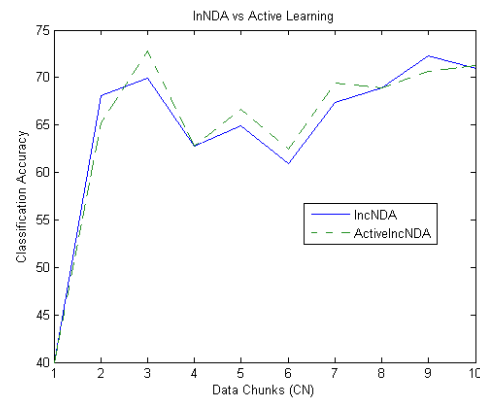


Figure A.9: *D3 (Liver Disorder) 7NN*

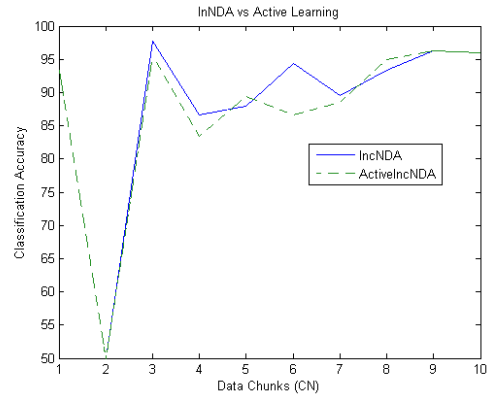


Figure A.10: D6 (Iris) 2NN

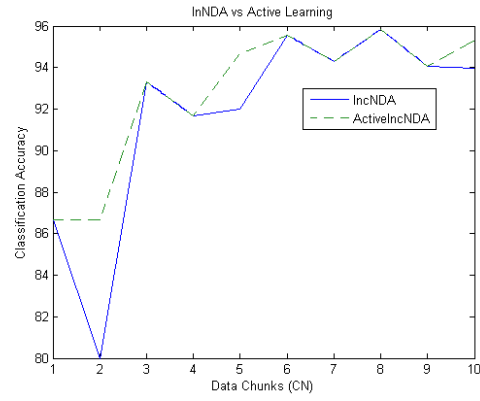


Figure A.11: D6 (Iris) 5NN

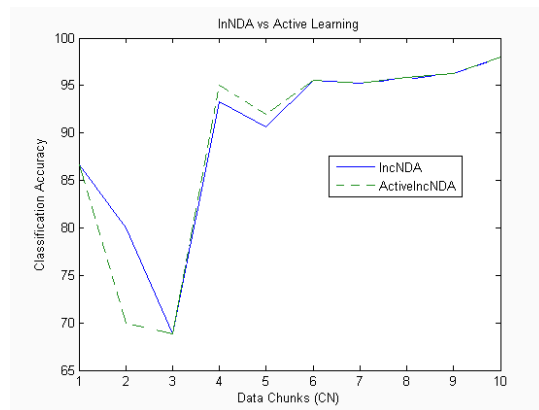


Figure A.12: D6 (Iris) 7NN

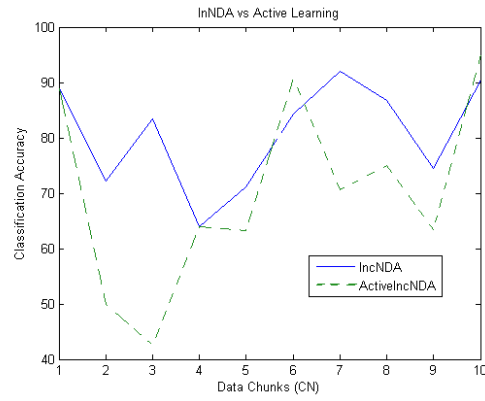


Figure A.13: D7 (Wine) 2NN

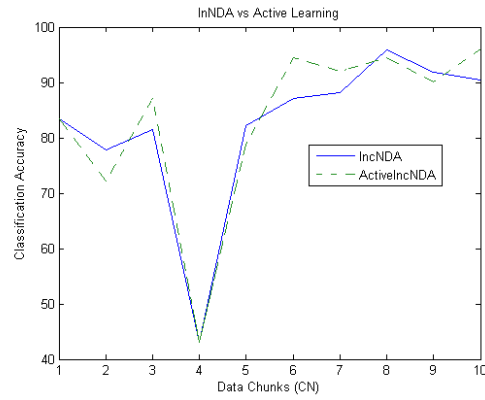


Figure A.14: D7 (Wine) 5NN

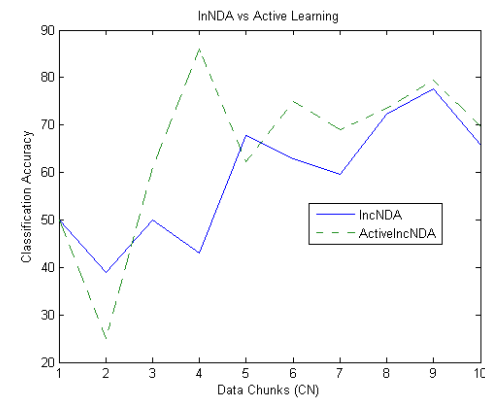


Figure A.15: D7 (Wine) 7NN

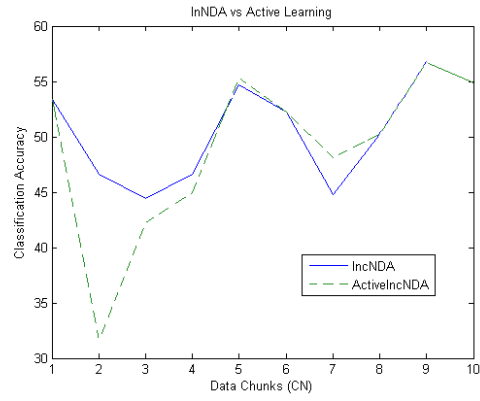


Figure A.16: D8 (Heart) 2NN

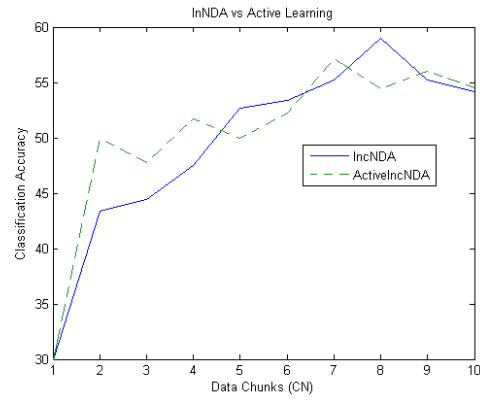


Figure A.17: D8 (Heart) 5NN

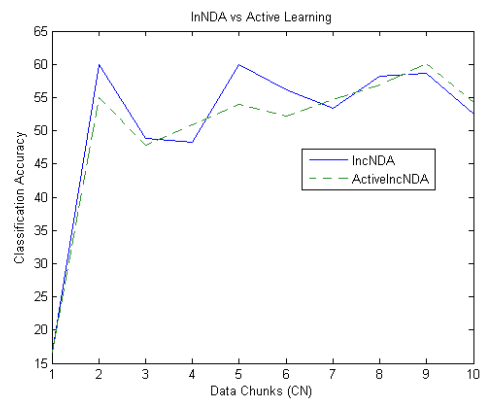


Figure A.18: D8 (Heart) 7NN

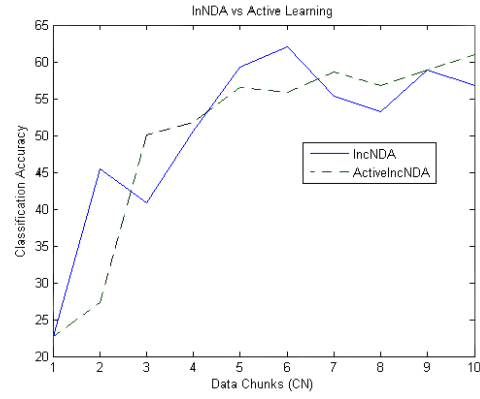


Figure A.19: D9 (Glass) 2NN

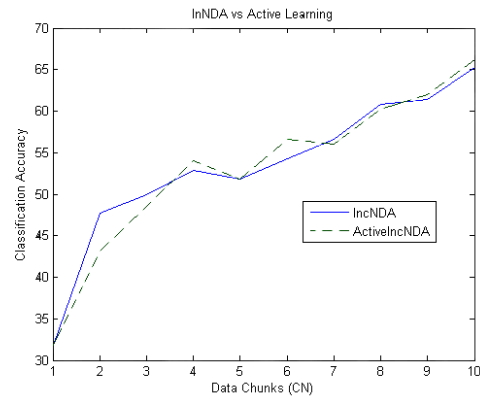


Figure A.20: D9 (Glass) 5NN

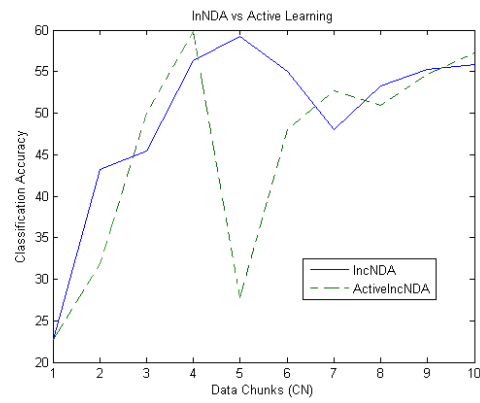


Figure A.21: D9 (Glass) 7NN

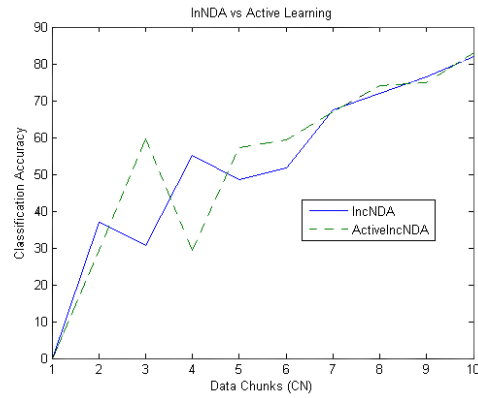


Figure A.22: *D12 (Face) 2NN*

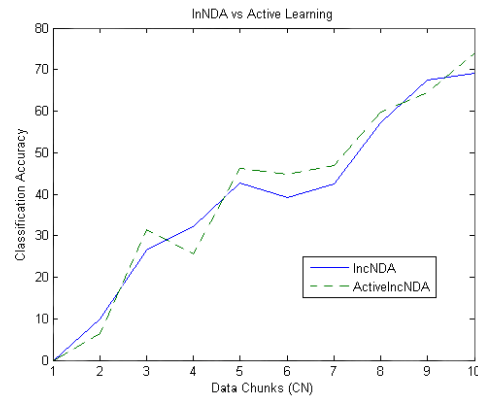


Figure A.23: *D12 (Face) 5NN*

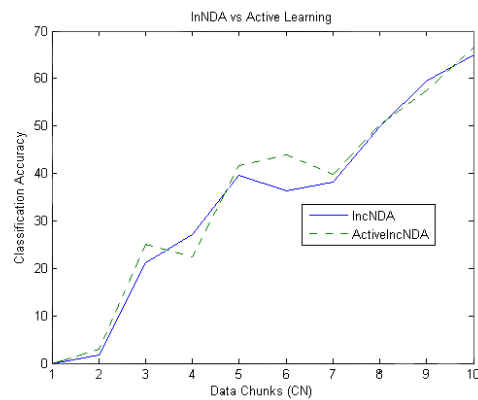


Figure A.24: *D12 (Face) 7NN*

Euclidean distance versus Pearson's product-moment correlation coefficient (PMCC):

For image datasets having affine distortion/transformation (as in ALOI dataset having wide baseline stereo object images), the conventional method (which uses Euclidean distance) for obtaining similar images is suboptimal, since, it is difficult to establish correspondence between stereo images by comparing regions of a fixed (Euclidean) shape, since their shape is not preserved under affine transformation. Therefore, we have used Pearsons product-moment correlation coefficient (PMCC) as a similarity metric. The comparison results of these two methods is shown in table A.1

Table A.1: Overall percentage accuracy comparison between Euclidean distance (*Euc.Dist.*) and Pearson’s product-moment correlation coefficient (*PMCC*) (similarity metrics) method, using Single-example and Multi-example method for ALOI datasets.

Method	Classification Accuracy %	
	PMCC	Euc.Dist.
Multi-example	73.53 ± 3.66	58.20 ± 5.57
Single-example	48.67 ± 22.26	44.97 ± 19.99

However, for face datasets (i.e. dataset 1 and dataset 2) the performance accuracy (%) was surprisingly the same for both PMCC and Euclidean distance method.

Appendix B

Conference Paper

Published in

'The Eighth International Conference on Information and Management Sciences'

July 20-28, 2009, Kunming & Banna, China.

Active Mode Incremental Nonparametric Discriminant Analysis Learning

Shaoning Pang, Kshitij Dhoble, Gary Chen, Nik Kasabov,
Tao Ban and Youki Kadobayashi

Abstract

This paper presents a novel active mode incremental nonparametric discriminant analysis (aIncNDA) learning method, in which previous passive incremental NDA is extended with data selective sampling, and performs active online discrimination analysis. Given an incoming instance y , the proposed aIncNDA computes a discrimination residue ratio between within-class and between-class ν , in which the residue is calculated using the k th regional nearest neighbor to class mean vector $\|NN_k(x, C) - \mu C\|$. The proposed aIncNDA is capable of estimating the discriminant contribution for every newly presented instance, because ν imitates the fundamental NDA $tr(S_b^{-1}S_w)$ criterion for a maximum separation between classes and minimum separation within classes. In the experiment, we described how the discriminative instances can be significantly selected based on discrimination residue with, at most, minor sacrifices in learning rate and classification accuracy. The experimental results show that the proposed aIncNDA performs gracefully under different level of redundancy, and the proposed aIncNDA learning system is capable of learning with less number of instances, but has more often an improved discrimination performance, than an passive incremental NDA.

Index Terms

Nonparametric Discriminant Analysis, Incremental NDA, Active Learning, Active Mode Incremental NDA Learning.

Shaoning Pang, Kshitij Dhoble, Gary Chen and Nik Kasabov are with Knowledge Engineering & Discovery Research Institute, Auckland University of Technology, Private Bag 92006, Auckland 1020, New Zealand Email: {spang, hhw9255, zh0202, nkasabov}@aut.ac.nz. Tao Ban and Youki Kadobayashi are with Information Security Research Center, National Institute of Information and Communications Technology, Tokyo, 184-8795 Japan Email: bantao@nict.go.jp, youki-k@is.aist-nara.ac.jp

I. INTRODUCTION

Active learning technique is crucial for classification as it iteratively selects distinctive information for training the classifier. Active rather than passive learning is preferred as it performs selective sampling, which enables the learning, immune to noise and data scarcity problems. Owing to its adaptive, evolving and dynamic characteristics it is potentially useful for targeted learning tasks and works well particularly for nonlinear dataset/data stream. By now, active learning has been successfully used in the field of internet security, bioinformatics [25] and text classification [14].

Active learning fundamentally consists of two main components namely the selective sampling engine and the base classifier. Selective sampling is carried out based on a certain criterion, which selects informative instances from the given chunk of data to better the learning function. Thus active learning technique is principally more accurate and computationally efficient than passive learning.

In supervised machine learning for class discrimination, the nonparametric discriminant analysis (NDA) is similar to Linear Discriminant Analysis (LDA) [21], which seeks a transformation towards a maximum separation between classes and minimum separation within classes. Classic NDA is a passive batch learning approach, assumes the entire dataset for training is truly informative and is presented in advance. However in real world applications, data is often being presented at different times in a stream of random chunks, and the quality of data is often not guaranteed due to noise affection. Incremental NDA (IncNDA) [19] somehow has solved the difficulty of NDA and empowered the NDA with an flexibility of incremental learning that accommodate a data stream sequentially. But in spite of that, IncNDA still conducts a rigid learning because IncNDA does not make any instance choices before actual learning, just passively learns whatever instances that are confronted/provided.

In order to overcome NDAs passive learning limitation, we have proposed an active mode incremental NDA learning approach, which incorporates incremental NDA (IncNDA) and selective sampling technique together to form an online active learning. The proposed aIncNDA allows constant informative update of NDA eigenspace obtained from the incoming data.

The rest of paper is structured as follows: Section 2 describes related researches and motivations. Section 3 introduce previous passive NDA learning approaches. Section 4 presents the proposed methodology been used in this experimentation. Section 5 contains comparative experimental results of IncNDA and aIncNDA. Finally in Section 6 conclusion is given along with future work directions.

II. RELATED RESEARCHES AND MOTIVATIONS

The concept of Active learning has only been explored recently. The key to active learning lies in its adaptive selective sampling technique, which selects the most informative instances or data, and eventually boosts the performance of the classifier. The selected data will be assimilated into the training set to retrain the classifier in order to achieve improved level of performance. This procedure can be

iterative, since the objective is to achieve a targeted level of performance with least amount of data and high number of informative instances. In our method, incremental NDA is addressed for active learning implementation.

A. Approaches of Active Learning

There are varieties of selective sampling approaches used in active learning models. Amongst them, one of the most commonly used technique is Pool-based active learning. However it suffers from multiple drawbacks. Most of the pool-based active learning iteratively selects random samples from the pool which may be informative or irrelevant [2]. Moreover, selecting the samples to be included in the pool itself is a time consuming process. Another selective sampling approach is membership query which selects samples directly from the dataset. Membership query scheme does not have the drawbacks posed by the pool-based scheme. It also reduces the predictive error rapidly and is less computationally intensive.

Clustering [6] and Batch mode active learning [7] are some of the other common flavors of active learning which aims at decreasing the redundancy amongst the selected instances, consequently providing more unique instances for the refinement of classifiers. Lastly, Query by Committee technique [8] is an effective approach, where selective sampling is based on the disagreement amongst ensemble of hypotheses. Some of the frequently used ensemble in this type of active learning includes techniques such as Bagging and Boosting.

For application, incorporation of active learning with support vector machine has been commonly used especially in the field of bioinformatics and text categorization [14]. However majority of them have made use of pool-based technique, which suffers from multiple drawbacks stated above, therefore it is recommended that though incorporation of active learning with SVM is good, other approaches such as membership querying or batch mode active learning should be used as they negate the drawbacks introduced by pool based learning.

B. Incremental Discriminant Analysis Approaches

It is well known that Linear Discriminant Analysis (LDA) [21] seeks a transformation towards a global maximum separation between classes and minimum separation within classes. In contrast, another known discriminant analysis approach, Nonparametric Discriminant Analysis (NDA) relies on local eigenvectors for obtaining discriminant knowledge from the entire dataset. The advantage of NDA over LDA is that, NDA does not rely on assumptions that instances are drawn from a given probability distribution, therefore are more robust than parametric methods such as LDA, and suits particularly on those nonlinear datasets. Similar to LDA, NDA requires the entire dataset for training presented in advance, thus is often called batch NDA in the literature. For incremental learning of NDA, Raducanu et.

al [19] proposed an incremental version of NDA, which allows us to maintain a constantly updated NDA eigenspace. However, both batch NDA and incremental NDA are merely a passive learning approach, learning passively whatever data is being given/confronted.

C. Motivation of Active Mode Incremental NDA Learning

To enable active learning of NDA, we incorporated incremental NDA and selective sampling technique together to form a new active learning technique, which delivers constant informative updating of NDA eigenspace, therefore minimizing concept drift and computational cost.

III. PASSIVE NDA LEARNING APPROACHES

Classic NDA [1] assumes that the entire training dataset is provided in advance, the learning is passively done in one batch. Incremental NDA (IncNDA) is capable of learning incoming instance continuously, but IncNDA also learns inactively whatever instances are confronted. The computation of Batch NDA and IncNDA are briefed as follows.

A. Nonparametric Discriminant analysis (Batch NDA)

Assuming that the data samples we have belong to N classes. Let C_i represents samples belonging to one of the class i , $i = 1, 2, 3, \dots, N$. Then, a NDA discrimination eigenspace according to [19] can be computed to express the class separability of data,

$$\Omega = tr(S_w^{-1} \cdot S_b) \quad (1)$$

In above Ω , S_w is the within class covariance matrix defined as:

$$S_w = \sum_{i=1}^{C_N} \sum_{j \in C_i} (x_j - \mu C_i)(x_j - \mu C_i)^T; \quad (2)$$

S_b is the between class covariance matrix defined as,

$$S_b = \sum_{i=1}^{C_N} \sum_{j=1, j \neq i}^{C_N} \sum_{q=1}^{n_{C_i}} W_{ijq} (x_q^i - \mu NN(x_q^i, C_j))(x_q^i - \mu NN(x_q^i, C_j))^T. \quad (3)$$

where μC_i is the mean vector of class C_i , and w_{C_i} is the number of samples in class C_i .

In S_b , $\mu NN(x_q^i, C_j)$ is defined as a local K-NN mean,

$$\mu NN(x_q^i, C_j) = \frac{1}{k} \sum_{t=1}^k NN_t(x_q^i, C_j), \quad (4)$$

where $NN_t(m_q^i, C_j)$ represents the t th nearest neighbor from vector m_q^i to class C_j . W_{ijq} is defined as a weighting function,

$$w_{ijq} = \frac{d^\alpha(x_q^i, NN_t(x_q^i, C_i))(x_q^i, NN_t(x_q^i, C_j))}{d^\alpha(x_q^i, NN_t(x_q^i, C_i)) + (x_q^i, NN_t(x_q^i, C_j))}. \quad (5)$$

where α denotes control parameter for sample weights which can be selected between zero and infinity.

B. Incremental Nonparametric Discriminant Analysis (IncNDA)

Consider new instances are presented in the future. Incremental NDA [19] incorporates the discriminant knowledge presented in the new coming sample as: given new instance y is coming in, then the current NDA model Ω is required to be updated as,

$$\Omega' = f(\Omega, y) = \text{tr}(S_w'^{-1} \cdot S_b') \quad (6)$$

This means that S_w and S_b are required to be updated respectively.

According to , the updated between class S_b' and within class S_w' covariance matrix can be calculated as follows:

$$S_b' = S_b - S_b^{in}(C_L) + S_b^{in}(C_{L'}) + S_b^{out}(y^{C_L}) \quad (7)$$

$$S_w' = \sum_{j=1, j \neq L}^{C_N} S_w(C_j) + S_w(C_L') \quad (8)$$

where $S_b^{in}(C_L)$ represents the covariance matrix between the existing class and the class newly presented, $S_b^{out}(y^{C_L})$ gives the covariance matrix between the existing class and the updated class $C_{L'}$, and $S_w(C_L')$ signifies the updated within class covariance matrix. For further computation approaches on $S_b^{in}(C_L)$, $S_b^{out}(y^{C_L})$, and $S_w(C_L')$, please refer to [19].

The above IncNDA can be used to construct an agent capable of updating the current discriminant knowledge $\Omega(t)$ by $\Omega(t+1) = \mathcal{F}(\Omega(t), \mathbf{y})$ whenever a new instance \mathbf{y} is confronted by the agent in the future. However, the IncNLDA is counted as a passive learning approach, because the IncNDA learns passively every instance confronted, even if the instance is confirmed redundant or noise data.

IV. THE PROPOSED ACTIVE INCNDA (aINCNDA)

For active learning, we consider here an active learning way (aIncNDA) to empower the IncNDA with the ability of detecting the discriminative interestingness of data before it is delivered for IncNDA learning. That is, the above IncNDA can be renovated to conduct incremental learning in an active learning way,

$$\Omega(t+1) = \begin{cases} \mathcal{F}_c(\Omega(t), \mathbf{y}) & \text{if } L(t) > \xi \\ \Omega(t) & \text{otherwise.} \end{cases} \quad (9)$$

where only discriminative instances are delivered for IncNDA learning. ξ is the threshold identifying discriminative criterion of NDA. The smaller ξ leads to the bigger number of instances learned by IncNDA.

Recall that the nature of NDA learning lies at the discriminability difference between the NDA transformed space and the original space. Straightforwardly, $L(t)$ can be represented as a type of mathematical residue that reflects the discriminability difference between the NDA transformed space and the original space.

Given one new instance presented at one time, similar to [20], the discriminability difference between the NDA transformed space and the original space of the IncNDA at time t by a classification performance evaluation as,

$$L(t) = Ad(t) - Ao(t), \quad (10)$$

where $Ad(\cdot)$ is the classification accuracy on discriminant eigenspace, and $Ao(\cdot)$ is the accuracy on original space. It could be any type of classification performance evaluation by any classifier.

However, such performance-based residue calculation involves a serious problem. That is, the $L(t)$ is highly classifier dependent. For example, suppose a K-NN method is used for performance evaluation $Ad(\cdot)$ and $Ao(\cdot)$, then the selected instances for incremental learning is meaningful only for K-NN classification and the category of prototype-based methods, but may not for the classification using any other methods such as hyperplane-based support vector machines (SVM) and decision-tree based C4.5.

A. Discrimination Residue Ratio

The idea of discrimination residue ratio is adapted from the weighting function (i.e. Eq. (5)) used in NDA, where $NN_k(x^i, C_i)$ and $NN_k(x^i, C_j)$ emphasize local within class distances and local between class distances. As we know, the principle of NDA, similar to LDA, seeks simultaneously minimizing within class distances and maximizing between class distances. The difference between NDA and LDA is, LDA is global model, whereas NDA focus on local instances distribution.

Given M new instances $Y = \{y_1, y_2, \dots, y_M\}$ presented as one chunk at time t , for each instance $y_i \in Y$, we can quickly estimate the within-class residue to the class mean vector μC_i :

$$\|NN_k(y^i, C_i) - \mu C_i\|, \quad (11)$$

also the between-class residue to any other the class mean vector $\mu C_j, j = 1, \dots, C_N, j \neq i$:

$$\|NN_k(y^i, C_j) - \mu C_j\|. \quad (12)$$

Thus, the contribution of incoming instance y_i to the NDA fundamental maximum $tr(S_w^{-1} S_b)$ criterion can be estimated as the following discrimination residue ratio of with-class to between-class scatter estimates

$$\nu(y_i) = \frac{\|NN_k(y^i, C_i) - \mu C_i\|}{\left\| \frac{1}{C_N - 1} \sum_{j=1, j \neq i}^{C_N} NN_k(y^i, C_j) - \mu C_j \right\|} \quad (13)$$

if $\nu(y_i) > 1$, then the contribution of y_i to NDA discrimination is positive, otherwise is negative.

However, it is noticeable that the above discrimination residue ratio varies in practice largely depending on individual dataset. Thus, it is hard for us to determine a suitable threshold value for a given dataset. To overcome this difficulty, we compute the discrimination residue ratio for every instance of the Y , then the above $\nu(y_i)$ can be normalized as,

$$\nu_{y_i} = \frac{\nu - \bar{\nu}}{\sqrt{\frac{1}{M} \sum_{m=1}^M (\nu_m - \bar{\nu})^2}} \quad (14)$$

where $\bar{\nu} = \frac{1}{M} \sum_{m=1}^M \nu_m$ is the chunk mean discrimination residue ratio. Thus, $L(t)$ in Eq. (15) can be implemented by ν_{y_i} as a chunk data filter.

$$\Omega' = \begin{cases} \mathcal{F}_c(\Omega, \mathbf{y}) & \text{if } \nu(y) > \xi \\ \Omega & \text{otherwise.} \end{cases} \quad (15)$$

V. EXPERIMENTS AND DISCUSSIONS

In this section, we have examined the efficiency and accuracy of the proposed aIncNDA method, and compared to IncNDA. Particularly, we investigate the relationship between 1) the discriminability and number of instances, 2) the redundancy and number of instances. To experiment on data with different discriminative characterization, we used datasets from two database resources. One resource is from UCI Machine Learning Repository [23], where we selected 8 datasets that have different application backgrounds and the features 100% of continuous/integer values and no missing value. The other resource is the MPEG-7 face database [24], which consists of *pose* and *light* two subsets, total 1355 face images of 271 persons, 5 different face images per person and each face image has the size of 56×46 .

A. Experimental Setup

To implement the proposed aIncNDA for incremental learning, we select randomly, for each dataset, 10% for initial batch NDA training, and divide the remaining data into 10 random chunks for incremental learning test. We collect every instance learned by aIncNDA, and evaluate the performance of aIncNDA and IncNDA on discrimination contribution at every learning stage. For performance evaluation, we compared the eigenspace from the proposed aIncNDA with the eigenspace from IncNDA by a leave-one-out kNN (k=1) classification over all data presented by current learning stage. Note that we use the term *learning stage* instead of the usual time scale since the events of data arriving in the above incremental learning may not happen in a regular time interval. Here, the number of learning stages is equivalent to the number of instances that have been learned by incremental models.

In the experiment, parameter ξ is relevant to the number of curiosity instances and the discriminability of the resulting NDA. For each experiments, we fixed ξ by the rule that the instances are significantly selected with, at most, minor sacrifices in discriminability.

B. Synthetic Dataset

We first experimented the proposed aIncNDA with a synthetic data set that has 3 classes 475 instances. The data distribution is a mixture of several 2D ([X1 X2]) Gaussian distributions as shown in Fig. 1.

Fig. 2 gives the distribution of the 257 informative instances learned by aIncNDA. As compared to the data distribution of the entire 475 instances, the discriminative representativeness of the selected

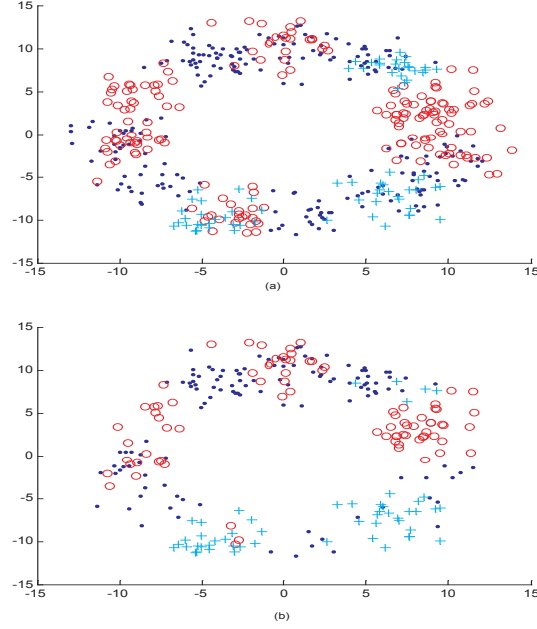


Fig. 1. The comparison of data distribution between the synthetic dataset and selected curiosity instances by proposed aIncNDA learning method. (a) the data distribution of the entire dataset; and (b) the data distribution of selected instance by aIncNDA.

instances by aIncNDA is clear because those 257 instances includes all critical instances for class distinction, such as instances involving class-mixture, and major representative instances of the independent class.

Fig. 2 illustrates the whole procedure of incremental learning with a comparison to IncNDA, where the horizontal and vertical axis represent the incremental stage and the classification accuracy from k-NN ($k=1$). As seen from the figure, the proposed aIncNDA and IncNDA is compared on the classification error at every incremental learning step. The classification accuracy difference between two methods is $+0.842105$, which indicates that the proposed aIncNDA achieves better learning effectiveness of the original IncNDA, although aIncNDA learns only 54.10% of total 475 instances.

C. UCI Datasets

Table I gives an comparison of aIncNDA versus IncNDA on the incremental learning of 8 UCI datasets. In the table, ξ is fixed for each dataset by the rule described above, the number of instances and the percentage to the number of all instances is denoted as ‘No. Instances(rate)’, and the classification accuracies is denoted as ‘Acc.’. The discriminability difference (denoted as ‘Diff.’) is calculated as the

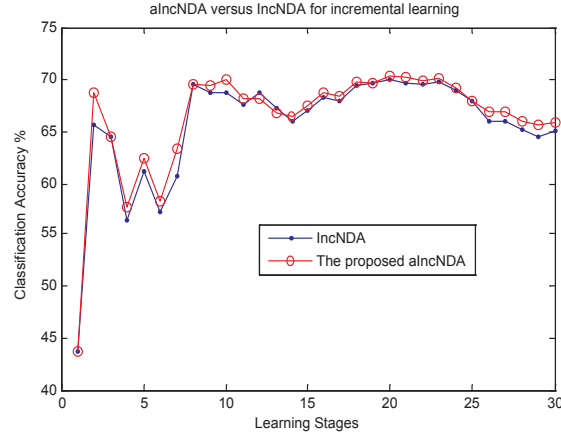


Fig. 2. The comparison of aIncNDA and IncNDA on the performance of incremental learning.

TABLE I

COMPARISON OF aINCNDA VERSUS INCNDA ON INCREMENTAL LEARNING OF INSTANCES OVER 8 UCI DATASETS.

Datasets	aIncNDA			IncNDA		Diff.[%]
	ξ	No. Instances(rate[%])	Acc.[%]	No. Instances	Acc.[%]	
Iris	0.75	56 (37.3)	94.5	150	92.0	+2.5
Liver-disorder	0.8	51 (22.2)	63.3	345	62.4	+0.9
Vehicle	3.0e-3	251 (29.7)	77.6	846	75.4	+2.2
Glass	0.98	50 (23.4)	60.1	214	52.5	+7.6
Wine	0.95	162 (92.7)	83.7	178	78.5	+5.2
Wisconsin	0.95	443 (95.7)	84.3	463	89.7	+1.1
Ionosphere	0.7	291 (83.1)	76.2	350	76.1	+0.1
Heart	0.65	33 (11.1)	53.2	297	52.3	+0.9

proposed aIncNDA minus IncLDA in terms of the K-NN LOO classification performance at the final learning stage.

As seen in the table, the proposed aIncLDA method, ignores 4.7%-88.9% instances of the whole dataset, constructs discriminant eigenspaces on the remaining 11.1%-95.3% selected instances. However, the discriminability of the obtained eigenspace from composed instance subset, compared to the eigenspace from all instances (using IncLDA), has no decrease, reversely, most of case has a slight increase. This suggests that the proposed active IncNDA learning is valid, and the selected instances by aIncNDA have the expected discriminative representativeness.

D. Performance under different discriminative redundancy

To test the performance of the proposed method under different level of discriminative redundancy, we carried out face recognition (FR) and face membership authentication (FMA) experiments [26], [27], [28] using the same face database described above. FMA is to distinguish the membership class (cls. 1) from the non-membership class (cls. 2) in a total group through a binary class classification. FMA involves more discriminative redundancy than face recognition problem, because the size of membership in FMA is often smaller than that of nonmembership, which indicates that not every instance are discriminatively important for FMA.

Over the 271 persons 1355 faces data, we conducted FR and FMA, respectively. For the FMA experiment, we set the membership size as 71 (cls. 1/cls. 2 is 71/200) without loss of generality. Thus, we compared the proposed aIncNDA with the IncNDA on incremental learning of 271 classes (i.e. FR) and 2 classes (i.e. FMA) data, respectively.

Fig. 3(a) shows the comparison of NDA discriminability between the proposed aIncNDA and the IncLDA for both FR and FMA experiments, and Fig. 3(b) reports corresponding the number of instances learned by aIncNDA.

As seen in Fig. 3(a), the proposed aIncNDA learns NDA for FR on 1331 of total 1355 instances, only 24 instances are found redundant. Whereas for FMA, aIncNDA learns 1093 of 1355 which is only about 20.0% of total 1355 instances are reduced. However, the performance of the proposed aIncNDA for both FR and FMA as given in Fig. 3(a) outperforms in most cases, the performance of the IncNDA on all 1355 instances. This indicates that the proposed aIncNDA is able to suit itself automatically to data with discriminative redundancy, and select a suitable number of instance to build an correct NDA model. This also can be reflect from Fig. 3(b), where aIncNDA is shown actively selecting different number of instance for incremental learning.

VI. CONCLUSION AND FUTURE WORKS

Method based on passive learning prove to be inadequate in real world application. To overcome this limitation, we have developed active mode incremental NDA which performs adaptive discriminant selection of instances for incremental NDA learning. Performance evaluation carried out on benchmark UCI datasets show that Active Mode Incremental NDA performs on par and in many cases better then incremental NDA with less number of instances. Given the nature of network data which is large, streaming, and constantly changing, we believe that our method can find practical application in the field of internet security.

Over the datasets from different resources, the proposed aIncNDA learning method is evaluated on: (1) aIncNDA versus IncNDA, and (2) performance under different level redundancy, where face recognition and face membership authentication are studied, respectively. The experimental results

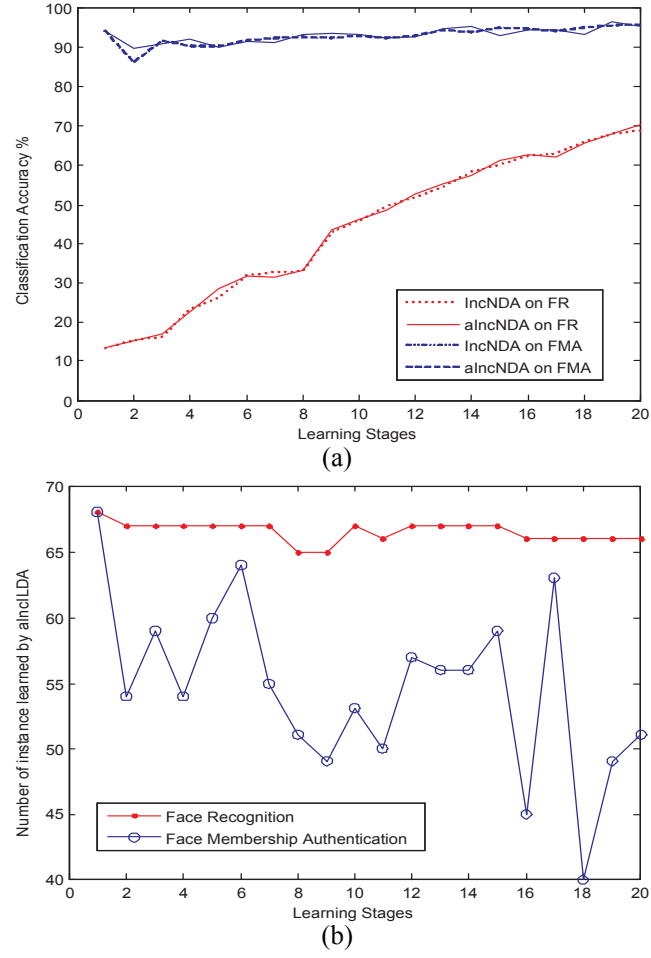


Fig. 3. Comparison of aIncNDA and IncNDA on FR and FMA, (a) the performance of aIncLDA versus IncNDA on incremental learning; (b) the number of learned instances by aIncNDA at every learning stage.

demonstrate that the proposed aIncNDA learning helps more efficient NDA learning with fewer instances, but with no performance deduction. One limitation of the proposed method concerns, as the original IncNDA retains raw data at every step of incremental learning, the data processing in aIncNDA is not one-pass.

As future work, the presented methods application in intrusion detection system will be exploited along with added enhancements to the selective sampling criterion. Also, the use of incremental classifier will be researched to serve as an extension to our present model which will eliminate the need for retraining further enhancing the processing speed while been computationally efficient.

ACKNOWLEDGEMENT

The authors would like thank Prof. Jie Yang for the useful discussion on this topic of discrimination analysis modelling during Dr. Pang's research visit at the institute of image processing and pattern recognition, Shanghai Jiao Tong University, in Dec. 2008.

REFERENCES

- [1] Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, Boston.
- [2] Charles, X. L. and D. Jun (2008). Active learning with direct query construction. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA, ACM.
- [3] Kai, Y., Z. Shenghuo, et al. (2008). Non-greedy active learning for text categorization using convex transductive experimental design. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. Singapore, Singapore, ACM.
- [4] Ming-yu, C., C. Michael, et al. (2005). Putting active learning into multimedia applications: dynamic definition and refinement of concept classifiers. Proceedings of the 13th annual ACM international conference on Multimedia. Hilton, Singapore, ACM.
- [5] Sanjoy, D. and H. Daniel (2008). Hierarchical sampling for active learning. Proceedings of the 25th international conference on Machine learning. Helsinki, Finland, ACM.
- [6] Hieu, T. N. and S. Arnold (2004). Active learning using pre-clustering. Proceedings of the twenty-first international conference on Machine learning. Banff, Alberta, Canada, ACM.
- [7] Steven, C. H. H., J. Rong, et al. (2006). Large-scale text categorization by batch mode active learning. Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland, ACM.
- [8] Prem, M. and J. M. Raymond (2004). Diverse ensembles for active learning. Proceedings of the twenty-first international conference on Machine learning. Banff, Alberta, Canada, ACM.
- [9] Danziger, S.A., Zeng, J., et al. (2007). "Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants." *Bioinformatics*, 23(13), 104-114.
- [10] Liu, Y. (2004). "Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification." *J. Chem. Inf. Comput. Sci.* 44(6): 1936 - 1941.
- [11] Warmuth, M. K., J. Liao, et al. (2003). "Active Learning with Support Vector Machines in the Drug Discovery Process." *J. Chem. Inf. Comput. Sci.* 43(2): 667 - 673.
- [12] Greiner, R., Grove, and D. Roth (2002). "Learning cost-sensitive active classifiers." *Artificial Intelligence*. 139(2):137-174.
- [13] Kapoor, A. and R. Greiner (2005). Learning and classifying under hard budgets.170-181.
- [14] Tong, S. and D. Koller (2001). "Support vector machine active learning with applications to text classification." *Journal of Machine Learning Research*, 2, 2001.
- [15] Yu, K. Bi, J. and V. Tresp (2006). "Active learning via transductive experimental design." In *International Conference on Machine Learning (ICML)*.
- [16] Zhang, W. V., X. He, B. Rey, and R. Jones (2007). "Query rewriting using active learning for sponsored search." In *ACM SIGIR Conference*.

-
- [17] Park, J. (2004) Convergence and application of online active sampling using orthogonal pillar vectors. *IEEE T. Pattern Anal. Mach. Learn.*, 28, 1197-1207.
 - [18] Bor-Chen, K. and D. A. Landgrebe (2004). "Nonparametric weighted feature extraction for classification." *Geoscience and Remote Sensing, IEEE Transactions on* 42(5): 1096-1105.
 - [19] Raducanu, B. and J. Vitria (2008). "Online nonparametric discriminant analysis for incremental subspace learning and recognition." *Pattern Analysis Application* 11: 259 - 268.
 - [20] Naoki Shimo, Shaoning Pang, Nikola Kasabov and Takeshi Yamakawa, "Curiosity-driven Multi-agent Competitive and Cooperative LDA Learning", *International Journal of Innovative Computing, Information and Control*, vol 4 n 7 pp 1537-1552, 2008
 - [21] P. Belhumeur, et al. "Eigenfaces vs. Fisher-faces: recognition using class specific linear projection", *IEEE Trans. on PAMI*, 19(7), pp. 711-720, 1997.
 - [22] Shaoning Pang, Seiichi Ozawa and Nik Kasabov, Incremental Linear Discriminant Analysis for Classification of Data Streams, *IEEE Trans. on System, Man, and Cybernetics-Part B*, 35(5), pp.905-914, 2005.
 - [23] <http://www.ics.uci.edu/mllearn/MLRepository.html>
 - [24] M. Kim, D. Kim, S. Bang, and S. Lee, Face Recognition Descriptor using the Embedded HMM with the 2nd-Order Block-Specific Eigenvectors, Jeju, Korea, ISO/IEC/ JTC1/SC21/WG11/M7997, 2002.
 - [25] S. A. Danziger, J. Zeng, et al., Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants, *Bioinformatics*, vol. 23, no. 13, pp. 104-114, 2007.
 - [26] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, and Sung Yang Bang, Constructing support vector machine ensemble, *Pattern Recognition*, vol. 36, no. 12, pp. 2757-2767.
 - [27] Shoning Pang, D. Kim and S. Y. Bang Membership authentication in the dynamic group by face classification using SVM ensemble, *Pattern Recognition Letters*, Vol. 24, pp. 215-225, 2003.
 - [28] Shaoning Pang, D. Kim, and S. Y. Bang, Face Membership Authentication Using SVM Classification Tree Generated by Membership-based LLE Data Partition, *IEEE Trans. on Neural Network*, Vol. 16, no. 2, pp. 436- 446 Mar. 2005.