การคัดเลือกลักษณะเฉพาะที่สามารถใช้ได้กับการแยกประเภทและการจัดกลุ่ม โดยการวิเคราะห์การแทรกสอดของแสงและค่าเอนโทรปีของการแยก

นายไพรสันต์ ผดุงเวียง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

FEATURE SELECTION APPLICABLE TO CLASSIFICATION AND CLUSTERING BY THE ANALYSIS OF OPTICAL DIFFRACTION AND ENTROPY SCORE DISCRIMINATION

Mr.Praisan Padungweang

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Doctor of Philosophy Program in Computer Science
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2011

Thesis Title	FEATURE SELECTION APPLICABLE TO CLASSIFICATION AND CLUSTERING BY THE ANALYSIS OF OPTICAL DIFFRACTION AND ENTROPY SCORE DISCRIMINATION
By	Mr.Praisan Padungweang
Field of Study	Computer Science
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.
Thesis Co-advisor	Khamron Sunat, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

…….. Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

…….Chairman
(Associate Professor Peraphon Sophatsathit, Ph.D.)

…….Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

…….Thesis Co-advisor
(Khamron Sunat, Ph.D.)

…….Examiner
(Suphakant Phimoltates, Ph.D.)

…….External Examiner
(Associate Professor Veera Boonjing, Ph.D.)

…….External Examiner
(Chularat Tanprasert, Ph.D.)

ไพรสันต์ ผดุงเวียง : การคัดเลือกลักษณะเฉพาะที่สามารถใช้ได้กับการแยกประเภทและการจัดกลุ่มโดยการวิเคราะห์การแทรกสอดของแสงและค่าเอนโทรปีของการแยก. (FEATURE SELECTION APPLICABLE TO CLASSIFICATION AND CLUSTERING BY THE ANALYSIS OF OPTICAL DIFFRACTION AND ENTROPY SCORE DISCRIMINATION) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: ศาสตราจารย์ ดร. ชิดชนก เหลือสินทรัพย์, อ. ที่ปรึกษาวิทยานิพนธ์ ร่วม: อาจารย์ ดร. คำรณ สุนัติ, 85 หน้า.

การรู้ลักษณะเฉพาะที่สำคัญแท้จริงของข้อมูลไม่เพียงแต่สามารถเพิ่มความเร็วในการเรียนรู้และการวิเคราะห์ข้อมูล แต่ยังทำให้ผลการเรียนรู้เหล่านั้นมีความถูกต้องมากยิ่งขึ้น ลักษณะเฉพาะที่สำคัญแท้จริงสามารถทำให้ผลการเรียนรู้มีความถูกต้องถึง 100% อย่างไรก็ตามการแยกลักษณะเฉพาะที่สำคัญนั้นไม่สามารถทำได้โดยง่าย ทั้งนี้เพราะว่าลักษณะเฉพาะที่เลือกมานั้นจะต้องคงไว้ซึ่งการกระจายตัวและโครงสร้างที่แท้จริงของข้อมูล ดังนั้นงานวิจัยนี้จึงได้นำเสนอวิธีการเลือกลักษณะเฉพาะแบบใหม่บนพื้นฐานของการจัดกลุ่มและการวัดแบบไม่มีผู้สอน โดยลักษณะเฉพาะจะถูกประเมินและเรียงลำดับใหม่ตามลำดับความสำคัญหรือเรียกอีกอย่างหนึ่งว่ากลวิธีการกรองลักษณะเฉพาะ ซึ่งเกิดจากการสังเกตการกระจายตัวของกลุ่มข้อมูลของแต่ละลักษณะเฉพาะ พบว่ามีความคล้ายกับการกระจายตัวของแสงที่ผ่านช่องเปิด การพิจารณาและประเมินลักษณะเฉพาะจึงเกิดจากแนวคิดของการเลี้ยวเบนของแสงผ่านช่องเปิด โดยใช้คุณสมบัติการแปลงฟูเรียร์ของความน่าจะเป็นของการกระจายตัวความหนาแน่น ซึ่งแนวคิดนี้เกิดจากสมมติฐานที่ว่าลักษณะเฉพาะที่มีการกระจายตัวของข้อมูลแสดงให้เห็นถึงการแยกสูงจะถือว่ามีความสำคัญ พร้อมทั้งได้พัฒนาให้สามารถประเมินลักษณะเฉพาะตามทิศทางการวางตัวของข้อมูล สิ่งที่ได้จากงานวิจัยครั้งนี้มีดังต่อไปนี้ (1) วิธีการใหม่ในการกรองลักษณะเฉพาะแบบไม่มีผู้สอนบนพื้นฐานของการวิเคราะห์การแยกเชิงแสง (2) วิธีการใหม่ในการให้คะแนนเทคนิคการกรองสำหรับการเลือกลักษณะเฉพาะที่ไม่มีการชี้แนะ และ (3) ความสามารถที่เหมาะกับการเลือกคุณสมบัติในงานประยุกต์เชิงแบ่งกลุ่มข้อมูลที่มีการชี้แนะและการจัดกลุ่มที่ไม่มีการชี้แนะ เมื่อเปรียบเทียบกับวิธี Laplacian score, SVD-Entropy และ LLDA-RFE ผลการทดลองแสดงให้เห็นถึงประสิทธิผลของวิธีที่นำเสนอ

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2554 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

4973886723: MAJOR COMPUTER SCIENCE

KEYWORDS: DISCRIMINATION ANALYSIS / UNSUPERVISED FEATURE SELECTION / FRAUNHOFER DIFFRACTION / FOURIER TRANSFORM / ENTROPY / PROBABILITY DENSITY ESTIMATION / INDEPENDENT COMPONENT ANALYSIS

PRAISAN PADUNGWEANG: FEATURE SELECTION APPLICABLE TO CLASSIFICATION AND CLUSTERING BY THE ANALYSIS OF OPTICAL DIFFRACTION AND ENTROPY SCORE DISCRIMINATION. ADVISOR: PROFESSOR CHIDCHANOK LURSINSAP, Ph.D., CO-ADVISOR: KHAMRON SUNAT, Ph.D., 85 pp.

Knowing the actual relevant features of a given data set not only can speed up the learning processes of classification or clustering algorithm, but also induce the higher prediction accuracy. Truly relevant selected features can make the prediction accuracy achieve 100%. However, it is not an easy task to distinguish the relevant features from the noisy features. This is because the selected relevant features must preserve the actual distribution and topological structure of the data space regardless of the original features. Therefore, a new feature selection based on unsupervised clustering and measure is proposed. The features are rearranged based on their relevant scores. This technique is called filter technique. Our approach is based on the observation that in any dimension, the distribution of clusters is similar to the scattering distribution of light passing through a set of vertical slits. The discrimination of data distribution is re-examined and evaluated using a simple observation motivated by the concept of optics diffraction. A property of the Fourier transform of probability density distribution is used. It is hypothesized that the features with high discrimination score are the relevant features. The criterion and algorithm are, then, extended to deal with data orientation whose direction of data alignment is defined by performing the discrimination evaluation on the bases locating towards the direction of data orientation. Then, the discrimination score of original features are computed. The key contributions from this research are: (1) new filter technique for unsupervised feature selection based on optical discrimination analysis, (2) new scoring of the filter technique for unsupervised feature selection, and (3) feasible capability to select features in both supervised classification and unsupervised clustering applications. Comparing with Laplacian score, SVD-Entropy, and LLDA-RFE, our experimental results show the efficacy of the proposed approach.

Department: Mathematics and Computer Science	**Student's Signature** .....
Field of Study: Computer Science	**Advisor's Signature** .....
Academic Year: 2011	**Co-advisor's Signature** .....

Acknowledgements

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, I would like to express my sincere gratitude to my advisor Professor Chidchanok Lursinsap for the continuous support and providing me with an excellent atmosphere for doing research. His guidance helped me all the time of research and writing of this dissertation. I would also like to thank my co-advisor Dr. Khamron Sunat who recommend me to studied in Ph.D. program and his invaluable assistance, helpful suggestions on research and study. Besides my advisors, I also express my thankfulness to my dissertation committee: Associate Professor Peraphon Sophatsathit, Dr. Suphakant Phimoltares, Associate Professor Veera Boonjing, and Dr. Chularat Tanprasert for their invaluable suggestions and for serving on my dissertation committee. I am grateful to Associate Professor Suchada Siripant for providing me with an excellent work environment in the Advanced Virtual and Intelligent Computing Center (AVIC) through out my study.

I would like to thank the Thailand Research Fund (TRF) for financial support of this research under the Royal Golden Jubilee (RGJ) Scholarship. This fund gave me a chance to pursue my Ph.D.

I greatly wish to thank Professor Chee-Hung Henry Chu for six months cooperation and helpful suggestions of being a visiting research scientist at University of Louisiana at Lafayette, Louisiana, U.S.A. To Professor Vijay Raghavan, Dr. Ryan Benton, and all members in LINC laboratory, thank you for sharing their knowledge and helping me during my stay. Also, I would like to extend my thanks to all Thai students at the Center for Advanced Computer Studies (CACS), University of Louisiana at Lafayette, for their kindness and cheerful support.

I cannot forget to thank my lab-mates in AVIC who have helped me in accomplishing this dissertation and all other academic activities.

I would like especially to thank my wife, Panida, for her unconditional understanding, extensive support, and standing beside me which nourishes me when I need it most. Finally, I would like to dedicate my Ph.D. to my family who have given me the opportunity of an education from the best institutions and support throughout my life.

Contents

Page

List of Tables

List of Figures

Figure Page

List of Abbreviations

Acronyms

DEODA	Discrimination Evaluation via Optic Diffraction Analysis
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
EVD	Eigenvalue Decomposition
ICA	Independent Component Analysis
FastICA	Fast Independent Component Analysis
ICs	Independent Components
LLDA-RFE	Laplacian linear discriminant analysis-based recursive feature elimination
LS	Laplacian score
SDF	Synthetic Discrimination function
NIR	Near Infrared
MLP	Multi-Layer Perceptron
kNN	k Nearest Neighbourhood
NMC	Nearest Mean Classifier
UCI	University of California at Irvine
WDBC	Wisconsin Diagnostic Breast Cancer
SRBCT	Small Round Blue Cells Tumor
ALL	Acute Lymphoblastic Leukemia
MLL	Mixed Lineage Leukemia
CH-index	Caliski-Harabasz index

List of Notations

Symbols

$\mathbf{X} \in \Re^{m \times n}$	A given data set which is a matrix of real number with m rows and n columns.
m	The number of feature or the number of dimension
n	The number of instance
$\mathbf{x}_j$	The j^{th} vector, i.e. the j^{th} instance, $\mathbf{x}_i = \{x_{1,j}, x_{2,j}, \ldots, x_{m,j}\}$
x_{ij}	The value of $\mathbf{X}$ at the i^{th} row of the j^{th} column, i.e. the i^{th} feature of the j^{th} instance
$\mathbf{v}_i$	The i^{th} feature vector $\mathbf{v}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$
$\bar{\mathbf{v}}_i$	the mean of the i^{th} feature vector
σ_i	The standard deviation of the i^{th} feature vector
LS_γ	The Laplacian score of the γ^{th} feature
$\dot{\mathbf{T}}$	The global similarity matrix
$\mathbf{T}$	The local similarity matrix
$\langle \mathbf{T} \rangle_{ii}$	The diagonal matrix of the summation of the similarity between the i^{th} and all instances
CE_i	The SVD-Entropy score value of the i^{th} feature
$\Phi(\mathbf{X})$	The entropy of the normalized eigenvalues (only positive values) of $\mathbf{X}^T\mathbf{X}$
τ_i	A singular values of a matrix
$\mathbf{\Psi}$	A projection matrix
$\dot{\mathbf{\Omega}}$	The global scatter matrix
$\mathbf{\Omega}$	The local scatter matrix
$\dot{\mathbf{L}}$	Normalized global Laplacian graph
$\mathbf{L}$	Normalized local Laplacian graph
ξ	The distance from reference line on the aperture plane
q	The distance from reference line on the observation plane
δ	Distance between slit
a	Silt width
λ	The wavelength
ω	The angular frequency
$A(\xi)$	The continuous version of aperture function
A_ξ	The discrete version of aperture function
$E(q)$	The disturbance at the position q on the observation plane (continuous version)
E_q	The disturbance at the position q on the observation plane (discrete version)
$P(q)$	The continuous version of intensity of the diffraction pattern
P_q	The discrete version of intensity of the diffraction pattern
$\Pi_a(\xi)$	A rectangular function with has width a and center at ξ
sinc(θ)	$\frac{\sin(\theta)}{\theta}$
G	Number of slit or number of cluster
Y_j	The height of the j^{th} principal maximum intensity
$f(t)$	The continuous version of function in time domain with time variable t
f_t	The discrete version of function in time domain with time variable t
$F(u)$	Fourier transform of function $f(t)$
F_u	Discrete Fourier transform of function f_t

List of Notations

Symbols

B	Number of evaluated frequencies (number of sampling)
H	The entropy which is used as the discrimination score of base
w_i	The weight value of the i^{th} histogram's bin
η	The bandwidth of the weight function
$\mathbf{M}$	The mixing matrix
$\mathbf{\Xi}$	The matrix of independent components (ICs)
$\mathbf{U}$	The matrix of basis vectors
$\mathbf{I}$	The identity matrix (matrix of standard basis vectors)
S_i	The discriminative score of the i^{th} feature
Q	The overall average performance over the base line
CH	The Caliski-Harabasz index (CH-index)
mCH	The modified Caliski-Harabasz index (mCH-index)

CHAPTER I

INTRODUCTION

Evaluation of relevant features of a given data set is one of the many important and necessary processes for data analysis. The relevant features can lead to achieve the high accuracy in supervised learning. Class label of instances, which can be viewed as an external knowledge, are used to evaluate the relevant features while the redundancy features are reduced using mutual information [1–5]. However, the results of unsupervised learning are provided by internal knowledge. Therefore, unsupervised learning can be applied to unclassified data. Moreover, unsupervised learning can alleviate the overfitting of an unreliable or mislabeled [6]. The selected feature subset is expected to help achieve better result on data analysis than use original features. But selecting the relevant features in unsupervised learning is more difficult than in supervised learning. This is because for the given unclassified data, there is no target associated with each training pattern in unsupervised learning. In addition, those selected relevant features must preserve the actual distribution and topological structure of the data space regardless of the original features. Since, a given set of data may consist of a mixture of noisy and relevant features. Those noisy features possibly affect the distribution and topology of the data space when the noisy data are projected to a lower dimensional space in the feature selection process. In general, finding the best feature subset which optimizes the defined criterion is an NP-hard problem. Since, finding the best feature subset requires an exponentially increasing (2^m) number of feature subset evaluations, which is in fact impractical if the data set has a large number of features. This also causes problem for supervised feature selection. Thus, good feature selection algorithm should have a good criterion and achieve small number of feature subset evaluations.

In real world situation the data sets tend to have a large number of features. However, if the number of variables is much larger than the number of data sample, it can cause problems in measurement [7, 8]. The computational cost of measurement will increase when the number of features increases. In addition, if there are noisy features, they can cause problems in classification or clustering results. Therefore, knowing the actual relevant features of a given data set not only can speed up the learning processes, but also can improve the classification and clustering accuracy. Discrimination analysis is one of interesting and popular approach to evaluate the rel-

evant features. It is one of requirement of both supervised and unsupervised learning. Because, there would be a high probability that the clusters are separated from each other for the selected feature subset with a high discrimination of density distribution. Therefore, this dissertation focus on filter methods for unsupervised feature selection based on the discrimination analysis.

1.1 Objectives

The main objectives of this research are the following:

1. To develop a new univariate filter technique for unsupervised feature selection.
2. To develop a new unsupervised feature selection based on a discrimination analysis.

1.2 Problem Statement

Given a data set which consist of unknown label of instances , we wish to rearrange features based on their discriminative value, and then, the feature subsets are selected from the ranking for evaluating.

1.3 Contribution

This dissertation proposed an unsupervised discrimination analysis based on physical optics principle. Fraunhofer approximation of optic diffraction was employed in the investigation of the density distribution. Firstly, the probability density distribution was treated as a synthetic aperture function. Then, the discrimination analysis was performed on a far-field diffraction pattern simulated by the Fourier transform of the aperture function. Finally, the discrimination evaluation was measured by using Entropy of magnitude at the middle of the bright areas. Moreover, the data orientation, which is direction of data alignment, was taken into account for evaluating the original features by performing discrimination evaluation on bases which are located toward a direction of data orientation. The algorithm was tested by setting up the experiments with benchmark data sets and compared the experimental results with the other existing methods using both classification and clustering algorithm.

1.4 Scopes of Work

In this research, the scopes of the work were constrained as follows:

1. The proposed technique was focused on filter technique for unsupervised selection.
2. The proposed filter technique was constrained by discrimination analysis.
3. The benchmark data sets were taken from UCI repository of machine learning database and publish database of Microarray data set.

1.5 Dissertation Outline

The rest of this dissertation is organized as follows: Chapter II is the related work and concept of proposed algorithm. In Chapter III, the backgrounds of optic diffraction pattern are briefly described. The discrimination analysis and the proposed algorithm are presented in Chapter IV. The implementations of the proposed algorithm and the experimental results on real world data set are in Chapter V. Chapter VI is the discussion and conclusion.

CHAPTER II

RELATED WORK AND CONCEPT OF PROPOSED ALGORITHM

2.1 Variation of Unsupervised Feature Selection

2.1.1 Type of Unsupervised Feature Selection according to its Evaluation

Unsupervised feature selection can be categorized, by its evaluation, into three techniques, filter technique, wrapper technique, and embedded technique. The filter technique selects the relevant features by looking only at the inherent properties of the data. In most cases, feature relevance score is individually calculated. A criterion for evaluating the quality of the features is defined, and then the features are rearranged according to its quality score. The variation of filter technique are univariate filter technique, collectively, and multivariate filter technique [9]. The wrapper technique evaluates the features by embedding the model hypothesis within the feature subset search. Finally, embedded techniques embeds the model hypothesis within specific classification or clustering algorithm. The features are simultaneously selected within the process of classification or clustering algorithm.

2.1.2 Type of Unsupervised Feature Selection according to its Characteristic

Unsupervised feature selection can be categorized, by its characteristic, into two main groups, 1) preserving of original properties of data or 2) discovery of required properties from data. The first group of feature selections is useful for reducing the computational time of algorithms. The example of properties which need to be preserved are variance and closeness between neighbourhood instances. The second group of unsupervised feature selections aim to increase the performance of clustering algorithms or classifiers. The example of required properties are data discrimination and multimodality density distribution. The evaluation step of both can be filter method or the dimension reduction technique. Other methods are the wrapper methods [10], and embedded methods [11–13]. However, filter methods can be used in flexible ways in the sense that they can be used as a data pre-processing without involving any classifier or

clustering algorithms. In the filter methods that follow the preservation approach, the features or dimensions, data space or transformed space, are ranked according to their preserving properties compared with the original data. This dissertation focus on filter methods for unsupervised feature selection which will be discussed in the next section.

2.2 Filler Method for Unsupervised Feature Selection

The categorized unsupervised feature selections by its characteristic are discussed in this section including preserving approach [14–17] and discovery approach [6, 18–20]. The preserving approach is useful for reducing the computational time of algorithms. A well-known and widely used technique is Principal Component Analysis (PCA). The given data are transformed into new orthogonal coordinates that are ranked according to the variance of the projected data. Classification or clustering algorithms can select a subset of the transformed features from the ranking for their evaluation purpose. Entropy of Singular Value Decomposition (SVD-Entropy) [14] used a greedy search strategy to select features in a multivariate manner, according to their preserving entropy of the Singular Value Decomposition (SVD) of the given data. Another interesting univariate technique is Laplacian score (LS) [15]. It is based on Laplacian Eigenmaps [21] and Locality preserving projections [22]. The feature scores are evaluated by their ability to preserve locality based on the observation of dissimilarity of neighbourhood instances in each feature and feature variance. The neighbourhood of each instance are measured by using the original data. A relevant feature according to the algorithm is a feature having the minimum value of the dissimilarity over the feature variance. However, as long as algorithms try to preserve the original data properties, they also preserve noise if the given data consist of noisy features.

The discovery approach aim to increase the performance of clustering algorithms or classifiers. A simple univariate technique is data variance in which features are ranked in decreasing order according to their variance values. The feature which has a larger value of variance is assumed to contain more information. An interesting and popular approach to evaluate the relevant features is discrimination analysis. A variety of discrimination analysis approaches are widely used and successfully applied in a supervised manner [23–25]. Closeness of instances belonging to the same class is preserved while the distances between different classes are maximized. Laplacian Linear Discrimination Analysis-based Recursive Feature Elimination (LLDA-REF) [6] is an unsupervised feature selection method that was proposed based on

a discriminative approach. Some proportion of Laplacian graph of the non-neighbourhood and neighbourhood data is re-examined as an evaluating function. The Laplacian graph of a neighbourhood as in the Laplacian score was used together with the Laplacian of a global graph such that all vertices are connected to each other. The feature ranking is produced in a multivariate manner based on recursive feature elimination. This algorithm gave an opportunity for discrimination analysis of unsupervised feature selection. However, there are some problems in using the discrimination analysis approaches in unsupervised data analysis. Although the neighbourhoods imply that the data may share some properties, e.g. they belong to the same cluster, we hardly expect the relationship of the non-neighbourhood data, i.e. using the global graph which is used in the LLDA-REF. They can belong to either the same cluster or the different clusters. The authors in [18] evaluated discriminative features that are evaluated using entropy and variance of Occurrence Numbers, a density distribution, which for one-dimensional data is a histogram without a bin with highest value. They imply that feature selection should be invariant, at least to some extent, with respect to metric scaling. However, histograms with arbitrary shape can have the same value of entropy and variance since the position of histogram bins are not included in the analysis. Therefore, in this dissertation, the shape of density distribution is investigated. The degree of overlap or discrimination of mode of density distribution was measured using a simple property of optics diffraction.

2.3 Related Work

The alternative methods for unsupervised feature selection were briefly reviewed in this section. Given a set of data samples $\mathbf{X} \in \Re^{m \times n}$, let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ and $\mathbf{x}_j = \{x_{1,j}, x_{2,j}, \ldots, x_{m,j}\}$. Each $x_{i,j}$ denotes the i^{th} feature of the j^{th} sample, $1 \leq i \leq m, 1 \leq j \leq n$. $\mathbf{X}$ can be view as a matrix of real numbers with m rows and n columns. The concept of the alternative methods are as follows:

2.3.1 Brief Concept of Laplacian Score (LS)

Laplacian Score [15] is a well known unsupervised feature selection based on filtering approach. It can identify features with larger variances as well as stronger locality preserving ability. The feature vector of $\mathbf{X}$ can be denoted by $\mathbf{v}_i$; $\mathbf{v}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$. Define $\bar{\mathbf{v}}_i = \frac{1}{n}\sum_{j=1}^{n} x_{i,j}$ as the mean value of the i^{th} feature. The Laplacian score of the γ^{th} feature, L_γ, which should be minimized and computed as follows.

$$LS_\gamma = \frac{\sum_{i=1,j=1}^{n} (x_{\gamma,i} - x_{\gamma,j})^2 T_{i,j}}{\sum_{i=1}^{n} (x_{\gamma,i} - \bar{\mathbf{v}}_i)^2 \langle \mathbf{T} \rangle_{i,i}} \tag{2.1}$$

where $T_{i,j}$ evaluates the similarity between the i^{th} and j^{th} samples and $\langle \mathbf{T} \rangle_{i,i}$ is an diagonal element of the diagonal matrix $\langle \mathbf{T} \rangle$. $T_{i,j}$ and $\langle \mathbf{T} \rangle$ are defined as follows

$$\begin{aligned} T_{i,j} &= \begin{cases} e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{t}} & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbours,} \\ 0 & \text{otherwise} \end{cases} \\ \langle \mathbf{T} \rangle_{i,j} &= \begin{cases} \sum_{r=1}^{n} T_{i,r}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

where t is a constant. $\mathbf{x}_i$ is the neighbour of $\mathbf{x}_j$ if the distance $||\mathbf{x}_i - \mathbf{x}_j||^2$ between $\mathbf{x}_i$ and $\mathbf{x}_j$ are ranked among the top k nearest distances measured from $\mathbf{x}_i$.

In equation (2.1), the dividend $\sum_{i,j}(x_{\gamma,i} - x_{\gamma,j})^2 T_{i,j}$ measures the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ only in the γ^{th} feature. The divisor $\sum_i (x_{\gamma,i} - \bar{\mathbf{v}}_i)^2 \langle \mathbf{T} \rangle_{i,i}$ is the variance used to normalized the dividend. The value of L_γ is minimum if the summation of the distances among all neighbours of the whole sample in the γ^{th} feature divided by feature variance is minimum. The relevant features indicated by Laplacian score are features with small value computed by (2.1).

2.3.2 Entropy of Singular Value Decomposition (SVD-Entropy)

SVD-entropy [14] is one of the preserving approaches. It ranks features according to the contribution of features. A feature is left out and, then the entropy of the singular value decomposition of the remaining features is computed. If there is nothing changed, compared with the entropy of the singular value decomposition of all features, then the feature is assumed to be an irrelevant feature. The contribution of the i^{th} feature to the entropy is defined as

$$CE_i = \Phi(\mathbf{X}_{[m \times n]}) - \Phi(\mathbf{X}_{[(m-1) \times n]}) \tag{2.2}$$

where $\mathbf{X}_{[(m-1)\times n]}$ is the removed i^{th} feature of the data matrix. $\Phi(\mathbf{X})$ is the entropy of the normalized eigenvalues (only positive values) of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{X}^T$ denotes matrix transpose. Let τ_i is a singular values of the matrix $\mathbf{X}^T$, then τ_i^2 is the eigenvalues of the matrix of $\mathbf{X}^T\mathbf{X}$. The entropy of the normalized eigenvalues can be computed as

$$\Phi(\mathbf{X}) = \frac{1}{m^{'}} \sum_{i=1}^{m^{'}} \tau_i^{'} \log \tau_i^{'} \tag{2.3}$$

where $\tau_i^{'} = \frac{\tau_i^2}{\sum_{j=1}^{m^{'}} \tau_j^2}$, and $m^{'}$ is number of the positive eigenvalues.

Any feature with high CE value is assumed to be the feature with high contribution. The SVD-Entropy ranked the features based on the CE value in a decreasing order.

2.3.3 Laplacian Linear Discriminant Analysis-based Recursive Feature Elimination (LLDA-RFE)

The LLDA-RFE [6] was extended from the Laplacian linear discriminant analysis (LLDA) to unsupervised cases. The LLDA algorithm aims to identify the features with high discrimination of instances between distinct classes. However, in the unsupervised case, the class label is not present. Therefore, the LLDA-RFE re-investigates the objective function of the LLDA algorithm for using in unsupervised feature selection. The LLDA-RFE computes feature score based on the projection matrix $\mathbf{\Psi}$ that maximizes the following criterion:

$$J_{LLDA}(\mathbf{\Psi}) = trace(\mathbf{\Psi}^T(\dot{\mathbf{\Omega}} - 2\mathbf{\Omega})\mathbf{\Psi}). \tag{2.4}$$

$\dot{\mathbf{\Omega}}$ and $\mathbf{\Omega}$ are the global and local scatter matrices. Let $\dot{\mathbf{T}}$, $\mathbf{T}$, $\dot{\mathbf{L}}$ and $\mathbf{L}$ be the global similarity matrix and the local similarity matrix, the normalized global, and local Laplacian matrices, respectively. These are defined as follows:

$$\begin{aligned}
\dot{\mathbf{\Omega}} &= \frac{1}{n}\mathbf{X}\dot{\mathbf{L}}\mathbf{X}^T, \\
\mathbf{\Omega} &= \frac{1}{n}\mathbf{X}\mathbf{L}\mathbf{X}^T, \\
\dot{\mathbf{L}} &= \mathbf{I} - \left\langle \dot{\mathbf{T}} \right\rangle^{-\frac{1}{2}} \dot{\mathbf{T}} \left\langle \dot{\mathbf{T}} \right\rangle^{-\frac{1}{2}}, \\
\mathbf{L} &= \mathbf{I} - \langle \mathbf{T} \rangle^{-\frac{1}{2}} \mathbf{T} \langle \mathbf{T} \rangle^{-\frac{1}{2}},
\end{aligned}$$

and

$$\dot{\mathbf{T}}_{i,j} = \begin{cases} e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{t}} & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

where $T_{i,j}$ evaluates the similarity between the i^{th} and j^{th} samples as in (2.2). $\left\langle \dot{\mathbf{T}} \right\rangle$ is a diagonal matrix defined as follows

$$\left\langle \dot{\mathbf{T}} \right\rangle_{i,j} = \begin{cases} \sum_r \dot{T}_{i,r}, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

Consider equation (2.4) $\mathbf{\Psi}$ can be found as the eigenvectors of $(\dot{\mathbf{\Omega}} - 2\mathbf{\Omega})$. The feature selection process is based on the recursive feature elimination. An eliminated feature implies that it is less relevant than the remaining features. Therefore, when the algorithm terminates, the features can be ranked in the relevant order. The LLDA-RFE algorithm is as follows:

Algorithm LLDA-RFE

---

input: no. of neighbourhood, data set $\mathbf{X}$ and no. of selected feature m^*.

output: data with selected features

1. Set $\vartheta \leftarrow m$.

While $\vartheta > m^*$ **Do**

2. Create the complete and k-nearest neighbour graphs on $\mathbf{X}$ and, then, compute $\dot{\mathbf{T}}$, $\mathbf{T}$, $\dot{\mathbf{L}}$ and $\mathbf{L}$;
3. Compute the SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{P\Lambda Q}^T$
4. Set $\mathbf{Z} = \mathbf{\Lambda Q}^T(\dot{\mathbf{L}} - 2\mathbf{L})\mathbf{Q\Lambda}$
5. Compute the EVD of $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{V\Delta V}^T$
6. Find the eigenvectors ($\mathbf{\Psi}$) of $\mathbf{PV}$ corresponding to the positive eigenvalues (Δ).
7. Remove the j^{th} feature with smallest score which is computed by
   $\sum_{i=1}^{m'} \sqrt{\Delta_i} |\Psi_{j,i}|$
   where m' is number of the positive eigenvalues.
8. Set $\vartheta \leftarrow \vartheta - 1$.

End While

Note that variables $\mathbf{P}, \mathbf{Q}, \mathbf{V}$ and $\mathbf{Z}$ are subject to re-defining in the next chapter.

2.4 Concept of the Proposed Algorithm

There are several possible approaches to evaluate the merit of each feature or a group of features. In this dissertation, the selection of relevant features is performed on each dimension based on the score computed by an evaluating function. The data are projected onto each dimension in the form of histograms. Suppose there are no overlaps among data clusters, each histogram can be used to represent each cluster and the envelope of the histogram can be viewed

as a light source, i.e. a slit. But if there are some overlaps among clusters, obviously the histograms are not clearly separated. The most relevant dimension should be the dimension that has the minimum amount of overlap between histograms. Hence, the problem of selecting the most relevant dimension or feature is transformed to a problem of measuring the degree of overlap of light sources. Figure 2.1 shows an example of projections and the envelopes of histogram of data clusters in a 2-dimensional space.

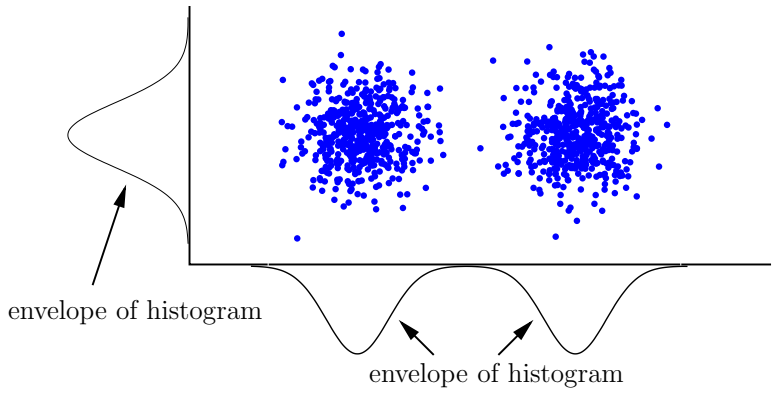

Figure 2.1: An example of projection of each data cluster onto each dimension and its corresponding envelope of histogram.

In the proposed approach, to measure the degree of overlap (discrimination), the set of envelopes is considered as a set of aperture slits and the far-field diffraction patterns of light penetrating through the slits and projected on a plane must be observed. By analyzing a well known aperture know as grating a connection between an intensity of diffraction patterns and the preferred properties was found. Based on this observation, the entropy of the intensity distribution of light is measured to evaluate the degree of overlap and to determine the relevancy of the corresponding feature. The diffraction is captured in terms of a Fourier transform. With this approach, it turns out that the proposed algorithm is invariant under feature scaling when directly deployed on the data space.

Light diffraction analysis and Fourier transform have been used in many fields. Casasent, Rozzi, and Fetterly [26] applied the Synthetic Discrimination function (SDF) to pattern recognition. Ostrovsky, Mota, and Cuatiaquiz [27] and Jing, Wong, and Zhang [28] used Fourier transform as a kernel function for pattern recognition in a supervised manner. Wu, Walczak, Penninckx, and Massart [29] used Fourier transform to obtain its coefficients for classifying the near infrared (NIR) data. The background of light diffraction and entropy measures are summarized in the next chapter.

CHAPTER III

BACKGROUND OF OPTIC DIFFRACTION PATTERN

The diffraction pattern can be explained by Huygens-Fresnel principle and its special case is based on the Fraunhofer approximation in terms of Fourier transform of an aperture function [30]. Suppose there is a plane wave of coherent light. A basic configuration for observing a diffraction pattern is shown in Figure 3.1. The aperture is assumed to lie on the left vertical plane. When the light travels through the aperture, its diffraction pattern occurs on the right plane called observation plane, which is at distance z and parallel to the aperture axis. The observation point is denoted by $\mathbf{o}$.

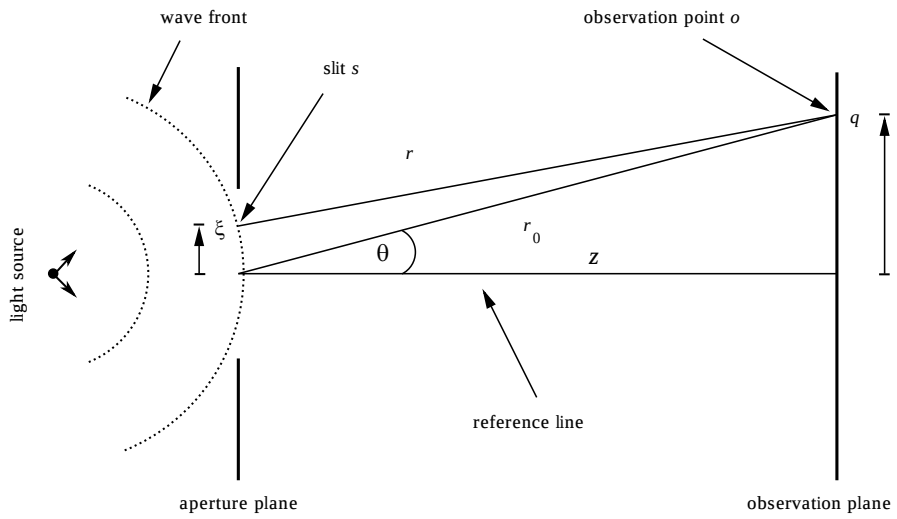

Figure 3.1: A basic configuration for observing a diffraction pattern. The left vertical solid line represents an aperture plane which can be viewed as a new light source while the right vertical solid line represents an observation plane.

The following notations will be adopted to distinguish a continuous variable from a discrete variable used in the function. Let V be any function with h as its variable. If h is a continuous variable within a given interval then function V is written as $V(h)$. Otherwise, it is denoted by V_h if h is a discrete variable.

3.1 Diffracted Wave under Huygens-Fresnel Principle

According to the Huygens[1]-Fresnel[2] principle, each point on the wave front can be regarded as a source of secondary wavelets radiating in phase with the same frequency as the original wave front. In Figure 3.1, the wave propagates onward from slit s to point $\mathbf{o}$ as a spherical wave front with radius of curvature equal to r. On the aperture plane at the considering position (ξ) from the reference line, let $A^{'}(\xi)$ be the electric field representation of light, and $A(\xi)$ be the aperture function. The relation between $A^{'}(\xi)$ and $A(\xi)$ is defined by the following equation.

$$A^{'}(\xi) = A(\xi)e^{i\omega t_0} \tag{3.1}$$

where $i = \sqrt{-1}$ is the imaginary value. ω is the angular frequency of the light, $\omega = 2\pi f$, and t_0 is an initial time. $A^{'}(\xi)$ is an complex value representation of light at the position ξ. For a plane wave at any given time, all positions on the wave front at the aperture are in phase, i.e. same phaser. Thus, $A^{'}(\xi)$ can be considered as a real value amplitude of light. The phase change of wave from slit $\mathbf{s}$ to point $\mathbf{o}$ depends on the distance r and the time t. By assuming the initial time is equal to zero and the amplitude of the light is constant along the propagation, the disturbance at $\mathbf{o}$ is, then, computed by

$$E^{(\xi)}(q) = A(\xi)e^{i(\phi r - \omega t)} \tag{3.2}$$

where $E^{(\xi)}(q)$ denote the disturbance at the considering position on the observation plane according to the light from the ξ position, $\phi = 2\pi/\lambda$ and λ is the wave length. Let $E(q)$ be a total disturbance a considering position at $\mathbf{o}$ with distance q from the reference line on the observation plane. The total disturbance at $\mathbf{o}$ is the summation of the contributions from all points on the aperture can be computed as

$$E(q) = e^{-i\omega t} \int_{\xi} A(\xi)e^{i\phi r} d\xi \tag{3.3}$$

where r is the distance between slit $\mathbf{s}$ and point $\mathbf{o}$. By changing the position of point $\mathbf{o}$, the diffraction pattern can be observed at any position on the observation plane. The intensity of the diffraction pattern, $P(q)$, is defined by the following equation

[1] Christiaan Huygens: 1629-1695. Dutch mathematician and physicist.

[2] Augustin-Jean Fresnel: 1788-1827. French physicist.

$$P(q) = ||E(q)||^2. \tag{3.4}$$

3.2 Fraunhofer Approximation

The Fraunhofer[3] diffraction, also known as the far-field diffraction, is based on the assumption that the wave front at the aperture is a plane wave and the diffracted waves are also plane. The observation plane is assumed to be located infinitely far away. This makes the diffraction angel, θ, in Figure 3.1 very small with respect to the reference line. Therefore, $\sin(\theta)$ can be approximated by $\sin(\theta) \approx \tan(\theta) \approx \theta$. Furthermore, the distance r can be estimated by

$$r = r_0 - \xi \sin(\theta) \tag{3.5}$$

where r_0 is the distance between the the position at the reference line on the aperture to the point **o** . Consequently, the total disturbance at point **o** as given by (3.3) can be rewritten as follows

$$E(q) = c \int_{\xi} A(\xi) e^{-2\pi i q \xi} d\xi \tag{3.6}$$

where $c = e^{i(\phi r_0 - \omega t)}$, $\phi = 2\pi\lambda$ and $q = \sin(\theta)/\lambda$. $A(\xi)$ denotes again the function of the aperture which is zero elsewhere outside the aperture. In general, $A(\xi)$ is equal to one where ξ is a position in the aperture. However, this function can be any arbitrary value, by allowing a real value of $A(\cdot)$ between zero and one. This can be regarded as transparent plates of various thickness. Any value of $A(\cdot)$ less than one means that the aperture absorbs the amplitude of light by allowing a fraction of the amplitude to transmit through the aperture. Hence, the function $A(\xi)$ can be viewed as an amplitude transmission function or an amplitude absorption aperture. It will be discussed again and be referred as a probability density distribution.

3.3 Example of Light Diffraction on well known Apertures

3.3.1 Double Slits Diffraction

Thomas Young[4] observed the diffraction pattern of light passing through two equal apertures called double slits, as shown in Figure 3.2. Suppose each slit has width a with center

[3] Joseph von Fraunhofer: 1787-1826. German optician and physicist.

[4] Thomas Young: 1773-1829. English physician and physicist.

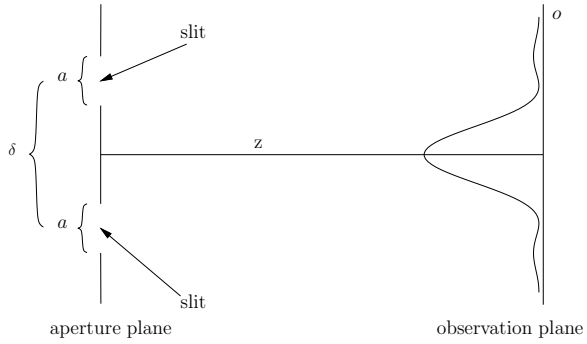

Figure 3.2: Double slits configuration for Young's observation with slits width a are placed with distance δ from each center. The thin solid line at the observation plane models the magnitude of diffraction pattern.

separated from each other by a distance of δ. The observation plane is at distance z and parallel to the aperture plane. The aperture function represents the double slits. The value of the function is equal to one only at the aperture and zero outside and it is defined as follows

$$A(\xi) = \Pi_a(\xi - \delta/2) + \Pi_a(\xi + \delta/2) \tag{3.7}$$

where

$$\Pi_a(\xi) = \begin{cases} 1 & \text{if } |\xi - a| \leq \frac{a}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

The $\Pi_a(\xi)$ denotes a rectangular function which has width a and center at ξ. The diffraction pattern observed by Fraunhofer approximation, using (3.6) is as follows

$$
\begin{aligned}
E(q) &= C \int_{-\frac{(\delta+a)}{2}}^{\frac{-(\delta+a)}{2}} e^{-2\pi i\varepsilon q} d\varepsilon + C \int_{-\frac{(\delta-a)}{2}}^{\frac{(\delta+a)}{2}} e^{-2\pi i\varepsilon q} d\varepsilon \\
&= C\left[-\frac{e^{-2\pi i\varepsilon q}}{2\pi iq} \Big|_{-\frac{(\delta+a)}{2}}^{-\frac{(\delta-a)}{2}} - \frac{e^{-2\pi i\varepsilon q}}{2\pi iq} \Big|_{\frac{(\delta+a)}{2}}^{\frac{(\delta-a)}{2}} \right] \\
&= \frac{-C}{2\pi iq}\left[e^{2\pi i\frac{(\delta-a)}{2}q} - e^{2\pi i\frac{(\delta+a)}{2}q} \right] + \frac{-C}{2\pi iq}\left[e^{-2\pi i\frac{(\delta+a)}{2}q} - e^{-2\pi i\frac{(\delta-a)}{2}q} \right] \\
&= \frac{-C}{2\pi iq}[\cos(\pi iq(\delta-a)) + i\sin(\pi iq(\delta-a)) - \cos(\pi iq(\delta+a)) \\
&\quad -i\sin(\pi iq(\delta+a)) + \cos(\pi iq(\delta+a)) - i\sin(\pi iq(\delta+a)) \\
&\quad -\cos(\pi iq(\delta-a)) + i\sin(\pi iq(\delta-a))] \\
&= -\frac{C}{\pi q}\left[\sin(\pi q(\delta-a)) - \sin(\pi q(\delta+a))\right] \\
&= -\frac{2C}{\pi q}\left[\cos(\pi q\delta)\sin(\pi qa)\right] \\
&= C2a\cos(\pi q\delta)\frac{\sin(\pi qa)}{\pi qa} \\
&= C2a\cos(\pi q\delta)\mathrm{sinc}(\pi qa)
\end{aligned}
\tag{3.9}
$$

where

$$
q = \frac{\sin(\theta)}{\lambda}, \quad \mathrm{sinc}(\theta) = \frac{\sin(\theta)}{\theta} \quad \text{and} \quad \mathrm{C} = \mathrm{e}^{-\mathrm{i}(\frac{2\pi z}{\lambda}+\omega \mathrm{t})}. \tag{3.10}
$$

The aperture is limited by range $\delta + a$. Suppose the limit of the aperture is fixed, i.e. $\delta + a$, to a *constant* and, then, change the distance of the slits. Increasing δ by $\Delta\delta$ means simultaneously reducing a by $\Delta\delta/2$. The magnitude of the diffraction pattern of two different values of slits distances are shown as Figure 3.3. The peaks of the magnitude which represent the center of the bright area are decreased when the observation position is far from the center position. Equation (3.9) can be regarded as an amplitude modulation of two signals. The magnitudes of the *cosine* are limited by the *sinc* function. If he height of the modes of magnitude are considered, i.e. the center of each bright area, those from Figure 3.3 (a) are more closely together than those from Figure 3.3 (b). In other words, the modes of magnitude from Figure 3.3 (b) are more Gaussian than the one from Figure 3.3 (a). This indicates that the equality of the modes can be viewed as the discrimination of slits. If the aperture is represented by means of a probability density distribution, the distance between modes of density distribution can be viewed as the distance between slits. A slit can be viewed as a data cluster. Clearly, when the peaks of diffraction

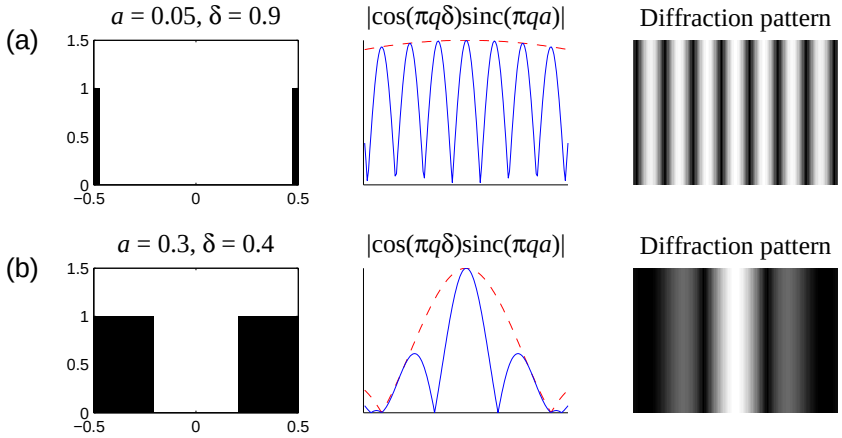

Figure 3.3: Double slits diffraction pattern using difference width of apertures. The top and the bottom rows illustrate, from left to right, the slits widths, the aperture functions, and the diffraction pattern.

pattern produced by this type of aperture are more closely together, it indicates that the clusters are more discrimination. These are also true for the intensity of the diffraction pattern as in (3.4).

3.3.2 Grating Diffraction

Let us now observe the diffraction pattern of light passing through a well-known aperture called grating. A grating consists of many of slits (grooves) per centimetre and is widely used in fibre optic communication systems [31].

Consider a grating with G similar slits shown in Figure 3.4 (a). Suppose the width of each slit is a with its center separated from each other by a distance of δ and $a \leq \delta$. If a is equal to δ, then the aperture become a single slit. The aperture function $A(\xi)$ represents the grating. The value of the function is equal to one only at the slit but zero outside and it is defined as follows

$$A(\xi) = \sum_{g=0}^{G-1} A^{(g)}(\xi) = \sum_{g=0}^{G-1} \Pi_a(\xi - \xi_g) \tag{3.11}$$

where $\Pi_a(\xi)$ denotes a rectangular function of width a and its center is at ξ as in (3.8). $A^{(g)}(\xi)$ is the g^{th} slit function. The diffraction pattern can be observed by Fraunhofer approximation by using (3.6). The intensity function of the grating can be expressed as in Appendix as follows

$$P(q) = C\left[\frac{\sin(G\pi\delta q)}{\sin(\pi\delta q)}\right]^2 \mathrm{sinc}^2(\pi a q) \tag{3.12}$$

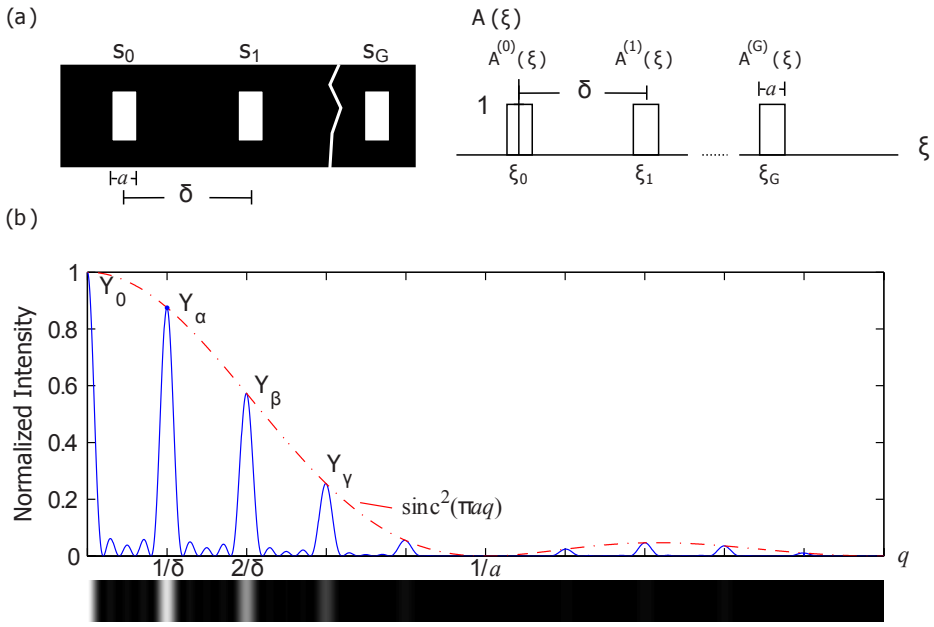

Figure 3.4: An example of grating and diffraction patterns. (a) The left image is a grating consisting of a group of G slits placed at the distance δ apart from each other. The width of each slit is equal to a. The right image is a set of corresponding $A^{(g)}(\xi)$ functions. (b) The normalized intensity of diffraction pattern. Only the range $q \geq 0$ is shown.

where $\mathrm{sinc}(\theta) = \frac{1}{\theta}\sin(\theta)$ and C is a constant.

Consider the effect of (3.12) on the diffraction pattern as shown in Figure 3.4 (b). Firstly, the first factor, i.e. the functions in the bracket, is based on the number of slits and the distance between slits. This factor produces patterns of bright areas that the principal maxima are located at $q = j/\delta$; $j = 0, \pm1, \pm2, \pm3, \ldots$. The principal maxima called zeroth-order and first-order principal maximum if $j = 0$ and $j = 1$ respectively. If the distance between slits increases, the distance between the bright areas of the diffraction pattern will decrease. The intensity of the bright areas next to the zeroth-order principal maxima will also increase, especially the first-order principal maxima denoted by Y_α in Figure 3.4 (b). Secondly, the last factor is based on the slit width, i.e. the $\mathrm{sinc}^2(\cdot)$, which is the dashed line in Figure 3.4 (b). It has a zeroth-order principal maxima at $q = 0$ and minima at $q = j/a$; $j = \pm1, \pm2, \pm3, \ldots$. It represents the effect of the single slit, which is the Fourier transform of (3.8). The grating with smaller slit width produces a diffraction pattern with more distance between the minima and the zeroth-order principal maxima. Also the value of function slowly decreases from one to zero in the interval $q = [0, 1/a]$. In addition, the values of the principal maxima of intensity are limited by this function as shown by the solid line in Figure 3.4 (b). Their positions can be approximated by $q = j/\delta$, thus, their values can be computed as

$$Y_j = \mathrm{sinc}^2(\pi\frac{ja}{\delta}), \tag{3.13}$$

that are increased if $\frac{a}{\delta}$ is decreased. Consequently, the principal maxima of intensity will increase closes to the zeroth-order principal maxima for both the large distance between slits and the small width of slits especially when $j = 1$, i.e. Y_α as shown in (3.13). Since its value decreases from one to zero when $\frac{a}{\delta}$ increases from zero to one. Therefore, if the equality the principal maxima of intensity were considered, it implies that the grating consists of a large distance between slits and small slit widths. Note that in this dissertation, a slit can be viewed as a cluster.

3.4 Discrete Fourier Transform of the Fraunhofer Diffraction

The Fraunhofer diffraction can be simulated by using discrete Fourier transform. In the frequency domain, The Fourier transform of a given function $f(t)$ with time variable t is a combination of all function values given by

$$F(u) = \int_{-\infty}^{\infty} f(t)e^{-2\pi iut}dt \tag{3.14}$$

where u denotes a frequency variable. To make $F(u)$ realizable, a discrete Fourier transform is considered. The value of $f(t)$ is sampled at consecutive time steps. Due to the fact that a discrete domain is being used, the following notation, f_t for $0 \leq t \leq N-1$, will be used to represent each sampled value at different time steps t.

The discrete Fourier transform (DFT) of each f_t, denoted by F_p for $0 \leq t, p \leq N-1$, is computed by the following equation:

$$F_p = \sum_{t=0}^{N-1} f_t e^{-\frac{2\pi ipt}{N}};\ p = 0, 1, \ldots, N-1 \tag{3.15}$$

However, in between two consecutive time steps, F_p can be made finer by allowing parameter p in the complex exponential power to be represented by $p = b(N-1)/(B-1)$, for $0 \leq b \leq B-1$. B be an arbitrarily chosen number of evaluated frequencies. Thus, the refined Fourier transform of F_p becomes

$$F_b = \sum_{t=0}^{N-1} f_t e^{-2\pi it\frac{b(N-1)}{N(B-1)}}, b = 0, 1, \ldots, B-1 \tag{3.16}$$

In (3.6), if the limit of the integral is extended to the range from $-\infty$ to ∞, there is no effect on the integral because $A(\xi)$ is zero outside the aperture. In addition, if the constant c in the integral is discarded, the total disturbance can be regarded as the Fourier transform of the aperture function. q can be viewed by means of the frequency of the Fourier transform in (3.14). Hence, the total disturbance can be observed based on the following relation

$$E_b \ \propto \ \sum_{\xi=0}^{N-1} A_\xi e^{-2\pi i\xi\frac{b(N-1)}{N(B-1)}},\ b = 0, 1, \ldots, B-1 \tag{3.17}$$

Note that if $b = 0$, then E_0 is the summation of all components of A_ξ. This is equivalent to the zero-angle coefficient, $\theta = 0$, of the total disturbance at the center of the observation plane in (3.6). Moreover, if the value of the aperture function is always real, then the total disturbance of the negative angle is the complex conjugate of that of the positive angle. Therefore, the intensity of both angles is exactly the same value. For those reasons, the first half of the total disturbance in (3.17) is needed, including the coefficient of position zero.

CHAPTER IV

DISCRIMINATION ANALYSIS AND PROPOSED ALGORITHMS

According to the optic diffraction principle mentioned in the previous chapter, the height of the principal maxima of the diffraction magnitude close to the zeroth-order principal maximum when the distance between slits is increased or the slits width are decreased. Consider the probability density distribution of 1-dimensional data. If the density of distinct clusters is considered as a slit and the distance between clusters is measured as the distance between slits, then the further distance between slits can be viewed as a greater discrimination of density distribution of the distinct clusters. The discrimination of density distribution via the diffraction principle and the feature discrimination analysis algorithm are proposed. The next sections will be discussed on the following issues with regards to the proposed algorithm:

1. Connection between optic diffraction and discrimination analysis.
2. Discrimination metric.
3. Aperture representation re-examined by means of the density distribution.
4. Base orientation and computation algorithm.
5. Feature evaluation algorithm.
6. Generalization of probability density distribution and basis orientation.

4.1 Connection between Optic Diffraction and Discrimination Analysis

In order to apply the principle of optic diffraction to feature evaluation, some constraints need to be considered. Suppose the limit of the aperture, i.e. $(G-1)\delta + a$, was normalized to a *constant* and, then, observe δ and a by analyzing the diffraction pattern. When sampling density distribution of each feature into the same discrete amount, e.g. the same amount of histogram bins, they can be viewed as the limited range aperture. The distance between modes of the

density distribution can be regarded as the distance between slits and each slit can be considered as a high density area. The discrimination of the high density areas is referred by means of the separation between slits, i.e. $\delta - a$.

Consider the value of principal maxima intensity that represents the center of the bright area next to the zeroth-order principal maximum intensity using a different grating as illustrated in Figure 4.1. The important factors subject to make the value of principal maxima close to each other, compared with Figure 4.1 (a), and their connection to discrimination analysis of density distribution are as follows

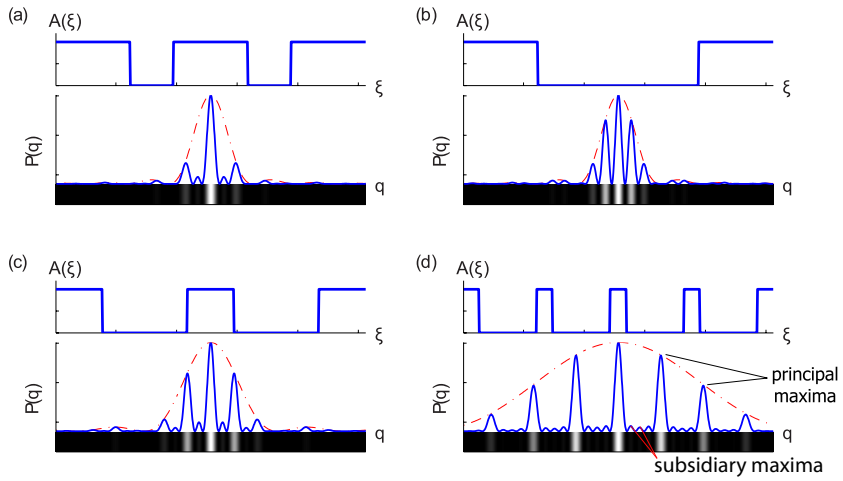

Figure 4.1: Comparison of the normalized intensity of diffraction pattern on different grating configurations.

- A larger δ causes the position of principal maxima close to the position $q = 0$, which is the highest value, and, thus, the principal maxima are increased. The number of slits also decreases because the range of the aperture is a constant. Figure 4.1 (b) illustrates the effect of larger distances among slits when compared to Figure 4.1 (a).

- A smaller slit width, a, with the same number of slits, G, causes the position of the first minimum of the *sinc* function (at $q = 1/a$ in Figure 3.4 (b)) to be located at a further distance from the center position and, thus, the principal maxima of the intensity are increased. The effect of slit width and the number of slits are shown in Figure 4.1 (a) and Figure 4.1 (c).

- Different numbers of slits cause different values of the principal maxima due to a/δ. The smaller a/δ makes the values of the principal maxima higher as discussed in Section 3.3.2.

This can be viewed as the higher density of individual cluster and larger discrimination of distinct clusters, as shown in Figure 4.1 (d) and when compared with Figure 4.1 (a).

Based on the observations above, the other density distributions can be analyzed as follows: First, if the silts function is a Gaussian function, i.e. $A_g(\xi) = e^{-\pi(\xi-\xi_g)^2/a^2}$, then the three issues above still hold for this type of distribution because the last factor in (3.12) is its normalized Fourier transform, which is $e^{-\pi q^2 a^2}$ instead of the $\mathrm{sinc}^2(\cdot)$. Therefore, in the interval $q = [0, 1/a]$, the smaller a using Gaussian function produces the similar effect to the principal maxima of intensity compared with one of the $\mathrm{sinc}^2(\cdot)$ functions. Next, if the density distribution is uniform, then the density has the same value for all positions. This implies that $A(\xi)$ is a rectangular function and the principal maxima of intensity are the principal maxima of the $sinc^2(\pi aq)$, where a is equal to the range of density distribution. When G is equal to one, the range of aperture is equal to a and the first factor of (3.12) is always equal to one. Finally, if the density distribution is a Gaussian distribution, i.e. $A(\xi) = e^{-\pi\xi^2/a^2}$, then there is only one maximum which is the zeroth-order principal maximum. This is because the Fourier transform of a Gaussian function is a Gaussian function consisting of only one maximum. Consequently, if the density distribution was observed by means of the aperture and the equality of the principal maxima of intensity, then the distribution can be ranked in the following order: (1) the most discriminative distribution, (2)uniform distribution, and (3) Gaussian distribution, respectively.

In general, $A(\xi)$ can be any arbitrary value where ξ is a position in the aperture. Suppose all positions on the wave front at the aperture are in phase and located at the same wave front. $A(\xi)$ can be considered as a real value amplitude of light. If the value of $A(\cdot)$ is between zero and one, it can be regarded as transparent plates of various thickness. Any value of $A(\cdot)$ less than one means that the aperture absorbs the amplitude of light by allowing a fraction of the amplitude to transmit through the aperture. Hence, the function $A(\xi)$ can be viewed as an amplitude transmission function or an amplitude absorption aperture. Figure 4.2 (a)-(d) show the diffraction pattern of other aperture configurations with a mixture of various height and width of Gaussian functions. The apertures are ranked according to a function of intensity maxima that will be proposed as a discrimination metric in the next section.

Consider the discrimination analysis from a Fourier transform point of view, then the observation plane of optic diffraction can be viewed as a frequency domain of the distribution function. The lowest frequency is at the center as the zeroth-order principal maxima intensity. Observing the height of the maxima can be viewed as observing the height of frequency of

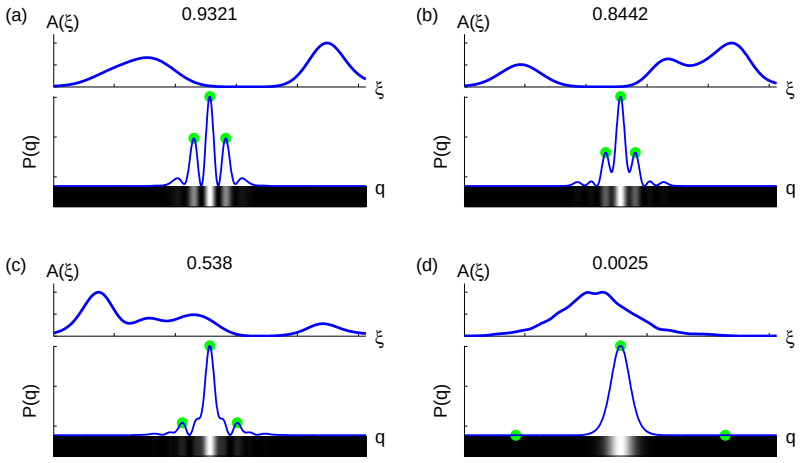

Figure 4.2: Comparison of the normalized intensity on different mixture of Gaussian functions. The first and second maxima are marked by the solid cycle.

the distribution. The optic analogy makes the observation clearer by analyzing the well known aperture function. Therefore, observing the height of the maxima means that high frequency exists in the frequency domain of density distribution. The high frequency means that the density rapidly changes from a high value to a low value and vice versa. This indicates that some separation exists between clusters.

4.2 Discrimination Metric

An evaluation using the diffraction pattern of an aperture represented by the probability density distribution is introduced in this section. Let H be a discriminatory score that is a function of the maxima intensity $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$; $Y_i \in [0, 1]$. The equality of $\mathbf{Y}$ can be easily measured by using the entropy and evaluated as follows

$$H = -\frac{1}{\log_\tau N} \sum_{i=1}^{N} Y_i' \log_\tau Y_i', \tag{4.1}$$

where τ is a constant. Normally, $\tau = 2$ and $Y_i' = Y_i / \sum_j Y_j$. Consider the equation (3.13), using the second maxima, $j = 1$, is enough for measuring the equality. Therefore, only the first two maxima of the mode of intensities for measuring the equality in (4.1), i.e. the zeroth-order principal maximum and the second maxima Y_α, are selected. The discriminatory score using the first two maximum modes of intensities Y_0, Y_α, using (4.1) is as follows

$$H = -\frac{1}{\log 2}\left(\frac{Y_0}{Y_0+Y_\alpha}\log\frac{Y_0}{Y_0+Y_\alpha} - \frac{Y_\alpha}{Y_0+Y_\alpha}\log\frac{Y_\alpha}{Y_0+Y_\alpha}\right). \tag{4.2}$$

Since Y_0 is the first maximum peak which is the coefficient E_0, Y_0 is a summation of the aperture function which is equal to one. Thus, the entropy of the normalized Y_0 and Y_α can be expressed by substituting $Y_0 = 1$ and $\log 2 = 1$ in (4.2) as follows

$$\begin{aligned}
H &= -\frac{1}{1+Y_\alpha}\log\frac{1}{1+Y_\alpha} - \frac{Y_\alpha}{1+Y_\alpha}\log\frac{Y_\alpha}{1+Y_\alpha} \\
&= -\frac{1}{(1+Y_\alpha)}\left[\log\frac{1}{(1+Y_\alpha)} + Y_\alpha\log\frac{Y_\alpha}{(1+Y_\alpha)}\right] \\
&= -\frac{1}{(1+Y_\alpha)}\left[\log 1 - \log(1+Y_\alpha) + Y_\alpha\log Y_\alpha - Y_\alpha\log(1+Y_\alpha)\right] \\
&= -\frac{1}{(1+Y_\alpha)}\left[Y_\alpha\log Y_\alpha - (1+Y_\alpha)\log(1+Y_\alpha)\right] \\
&= \log(1+Y_\alpha) - \frac{Y_\alpha}{1+Y_\alpha}\log Y_\alpha.
\end{aligned} \tag{4.3}$$

The largest value of (4.3) is equal to one when the value of the second maxima is equal to the value of the principal maxima. H can be viewed as a discriminatory score of density distribution. It can be viewed as a mapped value from Y_α to H as shown in Figure 4.3. This score value is used for basis evaluation and computing the discriminatory score of the original features.

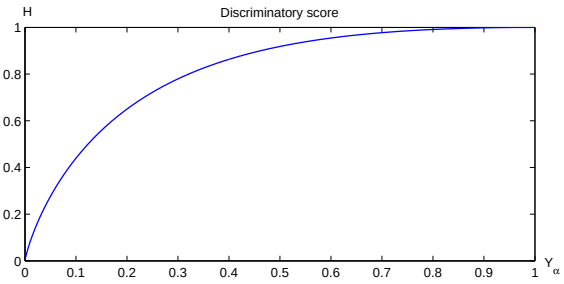

Figure 4.3: The discrimination score which is the entropy of the first two highest height of the principal maximum intensities.

4.3 Aperture Representation

In this section, the aperture function is referred to as a probability density distribution. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_j \in \Re^m$, be a data set. Each $\mathbf{x}_j = [x_{1,j}\ x_{2,j}\ \ldots\ x_{m,j}]^T$ is viewed as a column vector. Note that each $x_{i,j}$ can be considered as the i^{th} feature of data $\mathbf{x}_j$ Since our method considers one feature at a time, for any i^{th} feature, we form the i^{th} feature set, denoted by $\mathbf{v}_i$, from all data in $\mathbf{X}$ as a collection of $x_{i,j}$ for $1 \leq j \leq n$. Hence, we have $\mathbf{v}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$. Each feature set can be categorized into N bins of equal range and viewed as a density distribution of the data. This is known as a histogram. However, one of the difficulties in using this simple technique is choosing an appropriate width of bins which is sensitive to the outliers and the initial position. Therefore, the outliers were removed and a smooth version of the histogram was used instead. A low pass filter was applied to the histogram and can be viewed as a convolution between the value of histogram bins and a mask which is used to weight the neighbourhood bin values. For the i^{th} feature, the weight of the t^{th} bin compared with the considered r^{th} bin is defined as follows

$$\begin{aligned} w_i(r,t) &= e^{\frac{-(r-t)^2}{\eta_i^2}}, \\ \eta_i &= \frac{0.9(N-1)}{max_j(x_{i,j}) - min_j(x_{i,j})}\sigma_i n^{-1/5} \end{aligned} \tag{4.4}$$

where $r \in \{0, 1, \ldots, N-1\}$ are the indices of bins. η_i is adopted from [33] which is for multi-modal estimation. σ_i is the standard deviation of the given feature vector. Let μ_i be the mean of the i^{th} feature vector. Therefore, $\sigma_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{i,j} - \mu_i)^2}$. $x_{i,j} \in \mathbf{v}_i$ is an outlier if $|x_{i,j} - \mu_i| > 3\sigma_i$. The outliers are discarded for density estimation. The density estimation can be computed by **Algorithm 1**.

Algorithm 1: Density Estimation

input: a feature vector, $\mathbf{v}_i$, no. of bins, N.

output: probability density distribution with N bins

1. Remove the outliers.
2. Categorize the data into an array of N bins named A.
3. Apply a low pass filter
   $A_\xi = \sum_{t=0}^{N-1} A_t w_i(\xi, t)$
   where A_ξ is value of histogram of the ξ^{th} bin.
4. Normalize the histogram by
   $A_\xi \longleftarrow A_\xi / \sum_{t=0}^{N-1} A_t$

In **Algorithm 1**, the mean and the standard deviation are computed in step 1. Therefore, the time complexity of this algorithm is $O(n)$. This can be viewed as a special case of kernel density estimation, but much faster if the number of data is more than the number of bins. The density distribution that is represented by the histogram in array A is used as the synthetic amplitude absorption aperture or amplitude transmission function. The value of a bin indicates the remaining amplitude of light. The synthetic aperture completely absorbs the amplitude of light when the value of a bin is equal to zero.

4.4 Basis Orientation and Computation Algorithm

In a real world situation, the orientation of given data is normally unknown. The given data can lie in any direction which can cause data projection onto a standard basis (called an original feature) hard to analyze because distinct clusters can be projected onto the same position. Therefore, it is reasonable to explore the orientation of the given data before using **Algorithm 1**. The independent component Analysis (ICA) is a statistical method aiming to find the statistically independent component, not necessarily transformed by orthonormal bases. The algorithm can be divided into two main steps. First, the data are centered at the origin (zero-meaned) and uncorrelated (whitened). Second, some suitable optimization algorithms are used to find the independent components (ICs) with some constraints. FastICA is one of ICA algorithms maximizing the non-gaussianity distribution of the estimated components. This dissertation used the FastICA [34] algorithm to find the bases and orientation of the given data. Given a data set $\mathbf{X}$, it can be decomposed into two matrices as follows

$$\mathbf{X} = \mathbf{M}\boldsymbol{\Xi}, \tag{4.5}$$

where $\mathbf{M}$ is a mixing matrix and $\boldsymbol{\Xi}$ is a matrix of independent components (ICs). Both of them are unknown. However, (4.5) can be rewritten in the form of

$$\begin{aligned} \boldsymbol{\Xi} &= \mathbf{M}^{-1}\mathbf{X} \\ &= \mathbf{U}^{T}\mathbf{X}. \end{aligned} \tag{4.6}$$

If every vector in matrix $\mathbf{U}$ is constrained by a unit norm, then ICs can obviously be viewed as a linear combination of the given data and the bases. In addition, $\mathbf{u}_i \in \mathbf{U}$ can be viewed as a basis for transforming the input data into a new space. Therefore, an independent component consists of appropriate properties of original features constrained by a contrast function of the ICA algorithm. The univariate density estimation, as in **Algorithm 1** on an IC, can be viewed as the multivariate estimation on original features. Also, univariate evaluation on an IC can be referred to as an evaluation of a combination of original features. The bases can be computed as shown in **Algorithm 2**

The basis vectors in $\mathbf{U}$ are at the origin with the centred data $\mathbf{X}$. The ICs centred at the origin are the linear transformation of $\mathbf{X}$ using the de-whitened bases can be computed by (4.6).

Algorithm 2: Basis computation Algorithm

First step

1. center the data at the origin, $\mathbf{X}$

2. calculate eigenvalues and eigenvectors of the covariance matrix ($\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$) such that

$$\mathbf{CE} = \mathbf{ED}$$

where $\mathbf{E}$ is the matrix of orthonormal eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues

3. set $\mathbf{Z} = (\mathbf{D}^{-1/2}\mathbf{E}^T)\mathbf{X}$

where $\mathbf{D}^{-1/2} = diag(d_1^{-1/2}, d_2^{-1/2}, ..., d_{m'}^{-1/2})$ and m' is the numbers of positive eigenvalue

4. set $i = 1$

Second step

5. randomly choose an initial basis of unit norm, $\mathbf{u}_i$

6. if $i > 1$, orthogonalized basis by

$$\mathbf{u}_i \leftarrow \mathbf{u}_i - \sum_{j=1}^{i-1} (\mathbf{u}_i^T \mathbf{u}_j)\mathbf{u}_j$$

7. let $\mathbf{u}_i \leftarrow \mathbf{u}_i / ||\mathbf{u}_i||$.

8. let

$\mathbf{u}_o \leftarrow \mathbf{u}_i$ and

$\mathbf{u}_i \leftarrow \frac{1}{n}\mathbf{Z}(\mathbf{Z}^T\mathbf{u}_i)^3 - 3\mathbf{u}_i$

9. orthogonalized and normalized basis by steps 6, 7.

10. if $||\mathbf{u}_o - \mathbf{u}_i|| > \epsilon$ and $||\mathbf{u}_o + \mathbf{u}_i|| > \epsilon$ then

repeat steps 8 to 9.

11. if $i < m'$, set $i \leftarrow i + 1$, repeat steps 5 to 10.

12. de-whitened the basis, set $\mathbf{U} \leftarrow \mathbf{E}\mathbf{D}^{-1/2}\mathbf{U}$.

4.5 Proposed Feature Evaluation Algorithm

When the basis vectors are not standard bases, feature evaluation needs to be re-investigated. Suppose there is a unit vector $\mathbf{u}_j = [u_{1,j}, u_{2,j}]^T$ in $\Re^2$. The coordinates of each basis vector corresponding to the standard basis are $[1, 0]^T$ and $[0, 1]^T$. The cosine of the angle between a basis vector and its standard basis is the coefficient of the vector itself. For instance, if $\mathbf{u}_j = [1, 0]^T$ then the projection onto the first standard basis can exactly represent the vector $\mathbf{u}_j$, as shown in Figure 4.4.

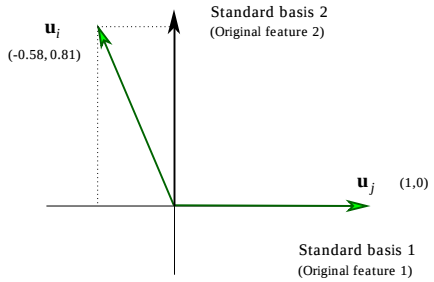

Figure 4.4: The unit vectors and the standard basis which represent the features. The vector $\mathbf{u}_i$ has a tendency to be parallel to feature 2 axis than feature 1 axis while the vector $\mathbf{u}_j$ is parallel to feature 1.

Therefore, the higher the value of the coefficient is, the more representative a standard basis can be. Selecting any feature as a discriminative representation will depend upon its discriminative score, denoted by S_i. The value of S_i is a weighted summation of coefficients from the set of basis vectors $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{m'}\} \in \Re^m$ and is defined as follows

$$S_i = \sum_{j=1}^{m'} H_j |u_{i,j}| \tag{4.7}$$

where H_j is the discriminatory score of basis vector $\mathbf{u}_j$. The process for evaluating relevant features in forms of basis vectors are summarized in the Discrimination Evaluation via Optic Diffraction Analysis (**DEODA**) algorithm.

DEODA Algorithm: Feature Evaluation Algorithm

Input: Matrix $\mathbf{U}$, all feature vectors $\mathbf{V}$, no. of samples B

Output: Set of discriminatory score for all features, $\mathbf{S}$

For each basis $\mathbf{u}_j$ **Do**

1. Project $\mathbf{V}$ with respect to $\mathbf{u}_j$

   $\mathbf{v}_j = \mathbf{u}_j^T \mathbf{V}$.

2. Estimate the probability density of $\mathbf{v}_j$ which represents the aperture by using Algorithm 1.

3. Compute the intensities of the far-field diffraction by applying the discrete Fourier transform

   $$P_b = || \sum_{\xi=0}^{N-1} A_\xi e^{-2\pi i \xi \frac{b(N-1)}{N(B-1)}} ||^2$$

   where $b = 0, 1, ..., B-1$ and $i = \sqrt{-1}$ is the imaginary value.

4. Find the second maximum mode of intensities, Y_α.

5. Compute the discriminatory score of the bases

   $H_j = \log(1 + Y_\alpha) - \frac{Y_\alpha}{1+Y_\alpha} \log Y_\alpha$.

EndFor

6. Compute discriminatory scores of original features

   $$S_i = \sum_{j=1}^{m'} H_j |u_{i,j}|, 1 \leq i \leq m$$

7. Rank features according to their discriminatory scores S_i in decreasing order.

The main computation of **DEODA Algorithm** is inside the *for loop*. Step 1 is the data projection, which is done in O(mn). In step 2, due to **Algorithm 1**, the time complexity is O(n). In steps 3 to 5, the computations are independent from the properties of data but are dependent upon the number of the histogram bins in step 3 which is a constant. Thus, the time complexity of these steps is O(1). Since these steps are iterated m' times, the overall time complexity of **DEODA Algorithm** is in O($m'mn$), where m is the number of features, m' is the number of bases, and n is the number of instances. However, if the bases are not standard bases, then some extra computation is needed to find the bases before apply the algorithm. If the bases are the standard bases, $\mathbf{U} = \mathbf{I}$, then the projection data in step 1 can be obtained directly from the data in each feature. Moreover, steps 5 and 6 need not be computed because the value of Y_α can be used as a discriminatory score of the original feature. The features are ranked in the same order. Consequently, the overall time complexity of **DEODA Algorithm** with $\mathbf{U} = \mathbf{I}$ is in O(mn).

4.6 Generalization

The main analysis of the proposed algorithm is performed on a probability density distribution of the projected data onto the bases. However, there have been open problems of generalization of both probability density distribution and basis orientation. In this dissertation, the simple estimation of the probability density distribution is proposed. The FastICA is deployed for discover the basis orientation, which is not necessarily orthonormal. Since the proposed probability density estimator is invariant to matrix scaling, thus, there is no need to pre-process the data by normalization in the case of $\mathbf{U} = \mathbf{I}$. An outlier is detected by comparing the distance between the corresponding feature vector and the mean of all feature vectors with three times the data standard deviation. Those methods can be changed arbitrarily according to the nature of the given data. Another concern is the number of sampled data for the Discrete Fourier transform. The fast Fourier transform algorithm [35] performs faster when the number of discrete values is a power of 2. This can be computed in advance before using the algorithm. In addition, the number of bins of the discrete probability density distribution should also be a power of 2 also.

CHAPTER V

EXPERIMENTAL RESULTS

5.1 Illustrative Examples by the Proposed Algorithms

The proposed algorithm was tested on a synthetic data. There were two experiments, namely evaluation on data space known as feature selection and on the transformed space known as feature extraction. Note that the proposed algorithm can evaluate features on data space using the information from the transformed space. This approach is also referred to as feature selection. All bases given in **DEODA Algorithm** are not necessarily standard bases. When the algorithm is applied to the bases from the ICA algorithm, the notation **DEODA** (ICs) is used to denote this process.

Consider the following example data in 20 dimensions or 20 features. The data set were generated with some noisy features so that they can affect the distance measure when all original features were used. The distribution of these data based on the first two features were shown in Figure 5.1 (a) for the feature selection experiments and in Figure 5.1 (b) for the feature extraction experiments. The data were smeared with different types of noise in the other remaining 18 features, namely uniform noise for features 3 to 8, Gaussian noise for features 9 to 14, and t-distribution PDF noise with degree of freedoms 1 to 6 for features 15 to 20. Prior to applying all algorithms, the data were normalized with zero mean and one unit variance for all features.

5.1.1 Feature Selection

Typically, in a feature selection approach, the features of the given data are rearranged according to their criterion. They are not restricted to any classification and clustering algorithm. One of the advantages of this approach is to reduce the search space from 2^m to m candidate subspaces. The feature subsets are subsets of sizes 1, 2, ..., m^*, ..., m, respectively, where feature subset of size m^* consists of m^* features having the highest score. It should be noted that although a univariate evaluation can reduce the search space, it can sometimes omit useful feature combinations. The classification and clustering algorithm evaluation are performed on the m subspaces.

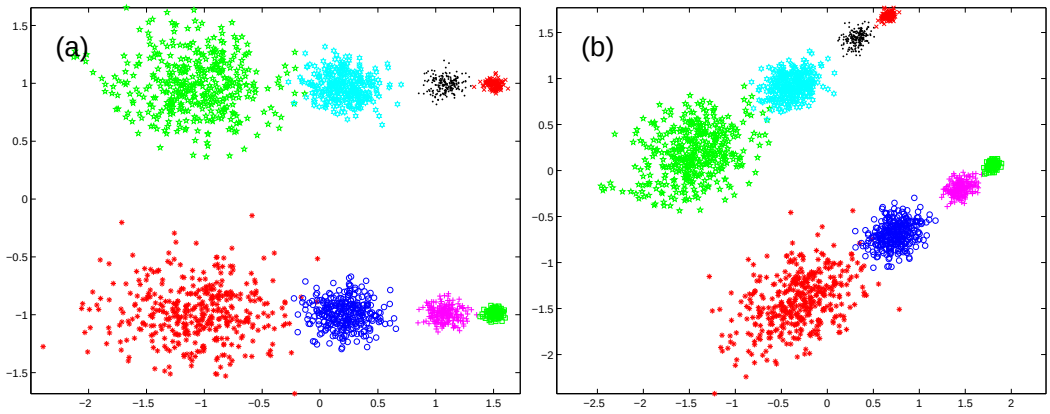

Figure 5.1: Synthetic data set. (a) The scatter plot of first two bases of original data. (b) The scatter plot of first two bases of transformed data.

In the example previously discussed, the probability density distributions of the projections onto the bases $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_5$, $\mathbf{u}_{10}$, and $\mathbf{u}_{15}$ of the first synthetic data set were shown in Figure 5.2 (a). The bases here are standard bases, $\mathbf{U} = \mathbf{I}$. The intensities of the diffraction and their patterns were shown in Figure 5.2 (b). Note that only the second maximum mode of the intensities was needed for computing the discriminatory score. The discriminatory scores of the corresponding bases were plotted in Figure 5.2 (c). To compare with the discriminatory scores of the other bases, Figure 5.3 shows the discriminatory scores of all bases. It can be seen that the first two bases had higher discriminatory scores than the others'. Thus, these two bases got the two highest ranks. The optimal classification and clustering algorithm needs to evaluate only the first two subspaces without noisy features.

The result was compared with the following measures: LLDA-RFE, SVD-Entropy, and Laplacian Score. Figure 5.4 shows only the first two features having highest scores. The proposed algorithm selected the features that showed clusters better than the features selected by the other measures.

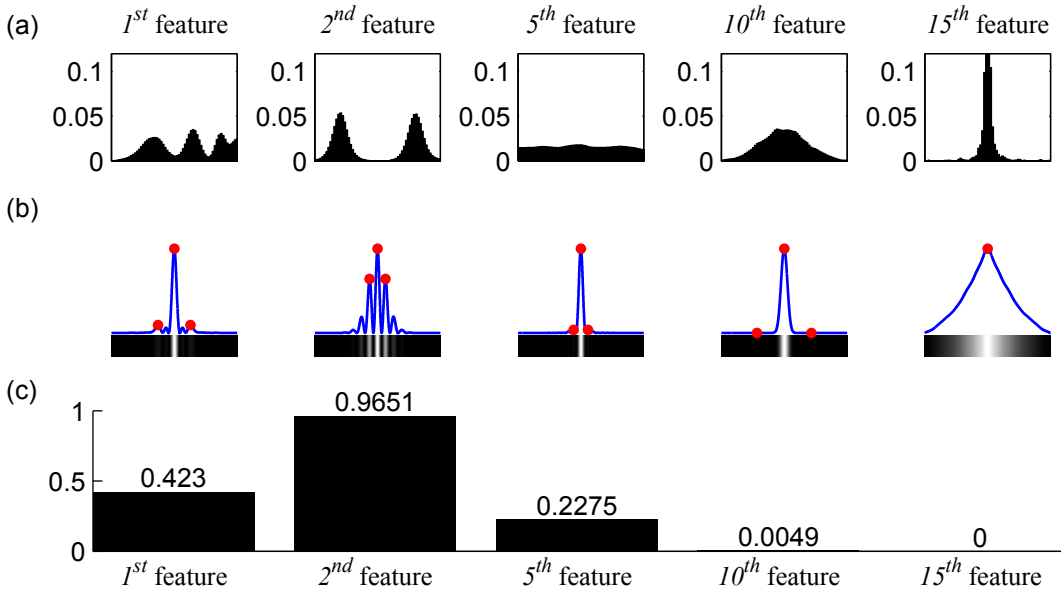

Figure 5.2: The density distribution, diffraction patterns, and discriminatory scores of bases 1, 2, 5, 10, and 15, respectively. (a) Density distribution of bases. (b) Diffraction patterns. (c) Discriminatory scores of bases.

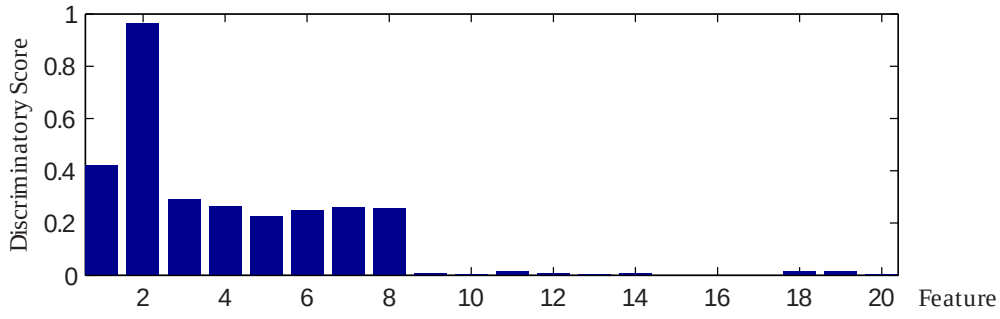

Figure 5.3: The discriminatory scores of all features of the first synthetic data set evaluated by the proposed algorithm.

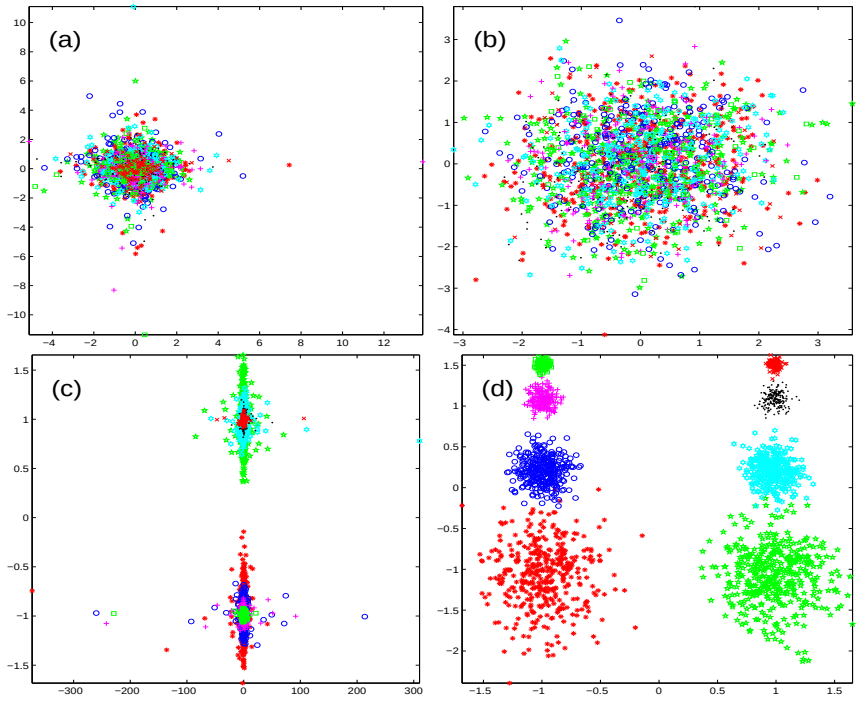

Figure 5.4: Scatter plots of the first two top score features evaluated by different algorithms. (a) LLDA-RFE. (b) SVD-Entropy. (c) Laplacian score. (d) The proposed algorithm.

5.1.2 Feature Extraction and Selection

To select the essential features, i.e. the relevant features in data space indicated by algorithms, the natural distribution direction, which is the direction of data alignment, must be computed first. This is followed by the measurement of the discriminatory score of each basis. PCA and ICA are considered to find the natural distribution direction of the given data and is here that a special technique of ICA called FastICA previously mentioned is used. The example in Figure 5.1 (b) was deployed here. Features containing different types of noise were added and, thus, produce only a few relevant features. For illustration purposes, first, PCA was applied to this data set and the bases with the highest eigenvalues, namely bases 1 and 2, were selected. Figure 5.5 (a_1) shows the distribution of data based on bases 1 and 2 having maximum eigenvalues. However, if the discriminatory score in equation (4.7) was used instead by setting H_i to the eigenvalues to select the essential features, the first two essential features became 19 and 18, respectively. The scatter plot of the data based on both features 19 and 18 was given in Figure 5.5 (a_2). Next, FastICA was deployed to the same data set. The same procedure was repeated with FastICA. Figure 5.5 (b_1) summarizes the distribution of data on the first two ICs in ICA

space and Figure 5.5 (b_2) shows the distribution of data based on the selected features 11 and 9, which were the first two highest scores computed by (4.7) by setting H_i to unity for all bases. In the case of DEODA algorithm, Figure 5.5 (c_1) and (c_2) show the distribution of data based on the bases with highest discriminatory scores after steps 5 and 6 in **DEODA Algorithm**. These were the first two highest scores of ICs and the first two highest scores of original features, respectively. The proposed algorithm can evaluate the true discriminatory features in both given space and transformed space.

Figure 5.6 shows another comparison of PCA, FastICA, and DEODA algorithm. The data are generated by randomly choosing some positions of an image as the first two features and, then, smeared them by some noisy features. Figure 5.6 (a_1) and (a_2) in the first column are the PCA results. Figure 5.6 (b_1) and (b_2) in the second column are the FastICA results. The last column is the results from the proposed algorithms.

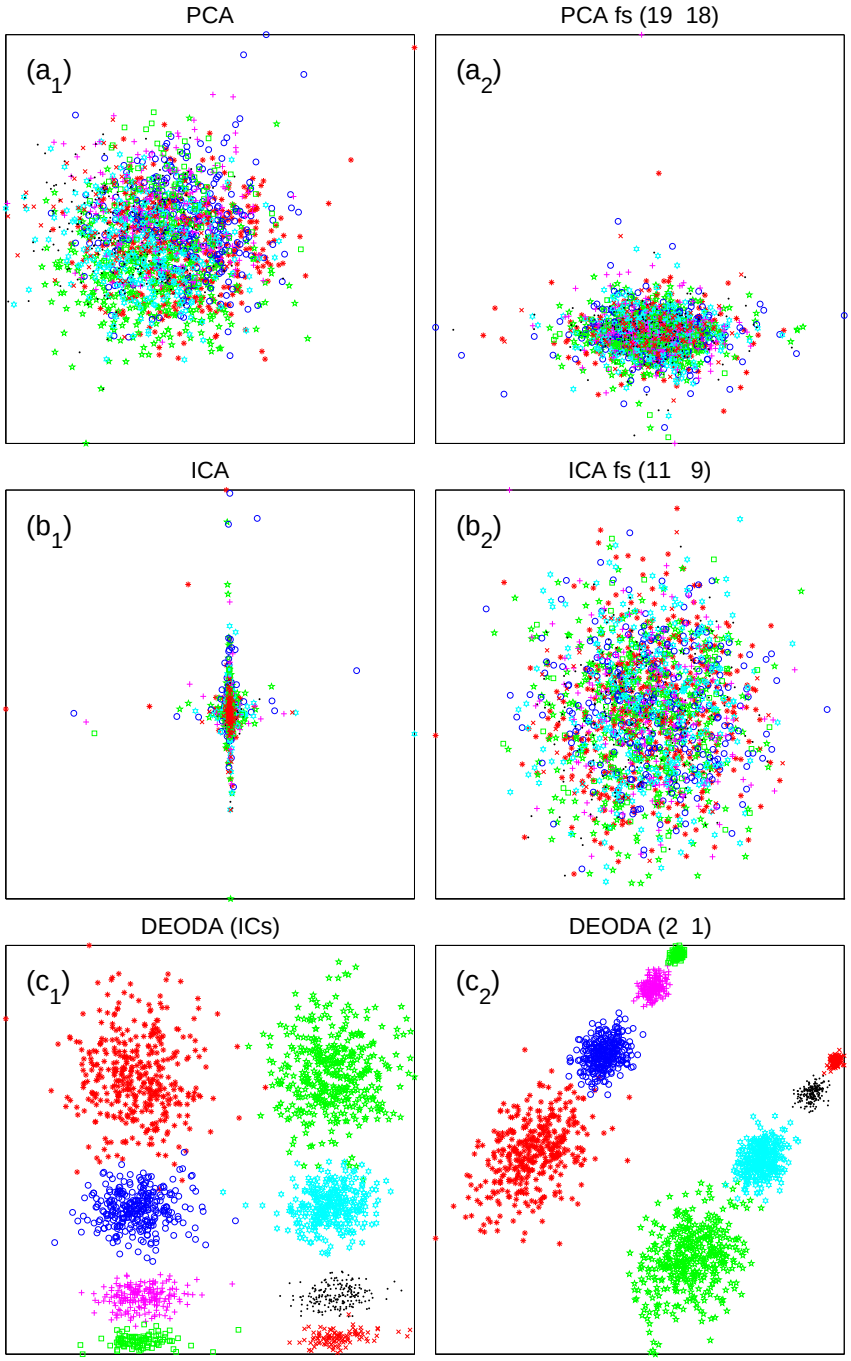

Figure 5.5: Scatter plots of the first two top scores of bases, (top row) in transformed space and (bottom row) features in given space using (4.7). (a) PCA. (b) FastICA. (c) DEODA.

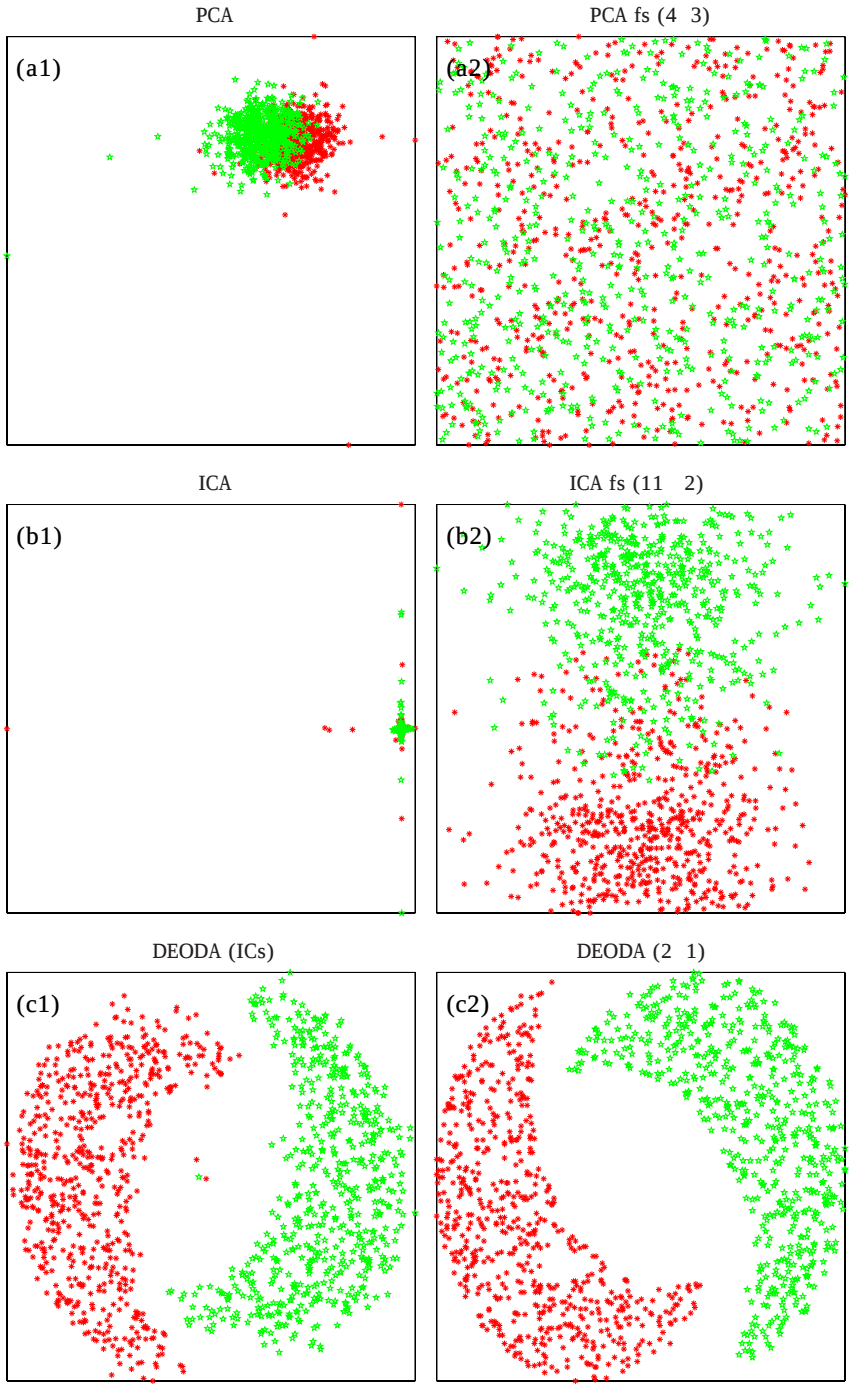

Figure 5.6: Scatter plots of the first two top scores of bases in transformed space (top figure) and features in given space (bottom figure). (a) PCA. (b) ICA. (c) DEODA.

5.2 Experiment Setting

The proposed algorithm is applied to classification and clustering tasks. In order to demonstrate the performance of the feature selection algorithms, data sets with less than 100 features from the UCI repository [36] as well as the data set with more than 2,000 features of gene micro-array data set were considered. The gene micro-array data set is one of challenging data set for feature selection, since it consists of a large amount of features but small number of samples. The features of the data sets are reordered in relevant order via the feature selection algorithm. Since our algorithm is not based on trial-and-error on all combinations of selected features, therefore, the evaluation was performed only on the m candidate subsets. The candidate subsets are subsets of sizes 1, 2, ..., m^*, ..., m, respectively, where feature subset of size m^* consists of m^* highest scored features.

The performance of our algorithm was compared with existing unsupervised algorithms including Laplacian score, LLDARFE and SVD-Entropy. All of them are unsupervised filter approaches that rearrange the order of the features of data based on the criterion of each algorithm. Class labels were not used in the ranking step of all algorithms. The experiments were performed in two different ways to compare the results from different spaces as follows.

In the first setting, all algorithms used information from the given data space to evaluate features. When it is in the case of $\mathbf{U} = \mathbf{I}$ and steps 1, 5 and 6 need no computation our algorithm will be named "DEODA". But, when the base are obtained from FastICA, the algorithm will be named "DEODA (ICs)". The number of bins, N, was set to 32 and the number of samples of DFT, B, was set to 1024 for all experiments as mentioned in section IV-F. For Laplacian score, the number of nearest neighbours for constructing the Laplacian graph was set to 5 as in [15]. For LLDA-RFE, the number of nearest neighbours for constructing the Laplacian graph was set to 3 as suggested in [6]. For SVD-Entropy, the features were evaluated based on the leave-one-out strategy as in [14].

In the second setting, all algorithms used information from transformed space to evaluate original features. The data sets were transformed into new spaces by the FastICA algorithm to find the new bases and ICs of data. In general, Laplacian score, SVD-Entropy and LLDA-RFE compute feature score using information from the given data space as the reference property. Therefore, when the data were transformed, the referenced property was computed using ICs instead. Then, the scores of the original features were computed as usual. Note that when the

SVD-Entropy was performed on the transformed space, the score values were added by the same constant and, therefore, the features were ranked in the same order. Our algorithm computed the scores of original features using the proposed strategy. The ICs were performed a priori and, then for each experiment, the same set of ICs was used.

5.3 Performance Measurement on Classification

This experiment was based on the assumption that a set of data belonging to the same class should be closely located to each others in the feature space and the selected features must preserve this assumption. To concern the closeness of data, the technique of k-nearest neighbour algorithm was used for evaluating the performance of the ranking. A 5-fold crossed validation strategy is used. Four subsets without class labels were used as the training set for feature ranking and the remainder were used as the test set for evaluating the performance of the algorithms. However, each random fold produced data sets with different distributions. Therefore, in each experiment, the data were iteratively and randomly divided into five subsets. There were 10 experiments and the averages of the accuracy were computed. For all experiments, all algorithms performed feature ranking on the same training set without class label and the feature subsets were evaluated on the same test set of data for comparison proposes.

Tables 5.1 - 5.8 show the average performance evaluations of UCI and micro-array data sets under different numbers of nearest neighbours. For micro-array data, the objective is to reduce number of genes causing diseases. Hence, each gene is considered as a feature and each disease is regarded as a class. For each percentage of accuracy, the superscript denotes the number of selected features and the subscript denotes the standard deviation of percentage of accuracy. For each data set, the highest performance obtained from different number of nearest neighbours were shown in bold numbers. The experimental results are summarized as follows.

Table 5.1: Average classification rate of selected feature subset using information from different spaces of the WDBC data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	91.52 $^{(17)}_{0.48}$	**92.91** $^{(19)}_{0.32}$	**92.94** $^{(17)}_{0.23}$	**93.28** $^{(19)}_{0.13}$	93.36 $^{(19)}_{0.09}$
SVD-Entropy	91.52 $^{(6)}_{0.41}$	92.89 $^{(7)}_{0.32}$	92.87 $^{(6)}_{0.17}$	93.24 $^{(7)}_{0.18}$	93.36 $^{(7)}_{0.09}$
LLDA-RFE	91.50 $^{(25)}_{0.52}$	92.89 $^{(30)}_{0.32}$	92.87 $^{(30)}_{0.26}$	93.24 $^{(30)}_{0.18}$	93.36 $^{(29)}_{0.09}$
DEODA	**91.60** $^{(18)}_{0.78}$	92.89 $^{(30)}_{0.32}$	92.87 $^{(30)}_{0.26}$	93.24 $^{(30)}_{0.18}$	93.36 $^{(30)}_{0.09}$
All 30 features	91.46 $_{0.38}$	92.89 $_{0.32}$	92.87 $_{0.26}$	93.24 $_{0.18}$	93.36 $_{0.09}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	91.46 $^{(30)}_{0.38}$	92.91 $^{(29)}_{0.37}$	92.91 $^{(12)}_{0.21}$	93.26 $^{(29)}_{0.15}$	93.40 $^{(19)}_{0.13}$
SVD-Entropy	91.52 $^{(6)}_{0.41}$	92.89 $^{(7)}_{0.32}$	92.87 $^{(6)}_{0.17}$	93.24 $^{(7)}_{0.18}$	93.36 $^{(7)}_{0.09}$
LLDA-RFE	91.46 $^{(30)}_{0.38}$	92.89 $^{(30)}_{0.32}$	92.87 $^{(30)}_{0.26}$	93.24 $^{(30)}_{0.18}$	93.36 $^{(30)}_{0.09}$
DEODA (ICs)	**93.82** $^{(27)}_{0.26}$	**94.26** $^{(27)}_{0.17}$	**94.61** $^{(27)}_{0.06}$	**94.78** $^{(27)}_{0.03}$	**94.54** $^{(27)}_{0.06}$
All 30 features	91.46 $_{0.38}$	92.89 $_{0.32}$	92.87 $_{0.26}$	93.24 $_{0.18}$	93.36 $_{0.09}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

A) WDBC Data Set consists of two classes of 569 data with 30 features. Tables 5.1 (A) and 5.1 (B) summarize the comparison results with each other method for both experiments. Laplacian score achieved slightly higher accuracy than those of other methods. DEODA algorithm showed the comparable mean results with those of the other methods for the given data space. However, for the transformed space, the selected feature subsets from DEODA achieved the highest average accuracy for all of number of nearest neighbours with smallest standard deviation.

Table 5.2: Average classification rate of selected feature subset using information from different spaces of the Sonar data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	82.06 $^{(43)}_{0.82}$	81.30 $^{(57)}_{1.67}$	78.72 $^{(55)}_{3.80}$	74.31 $^{(54)}_{3.09}$	69.42 $^{(51)}_{4.21}$
SVD-Entropy	**85.10** $^{(51)}_{2.21}$	**82.11** $^{(51)}_{3.28}$	**80.45** $^{(57)}_{3.96}$	76.99 $^{(46)}_{2.90}$	74.50 $^{(49)}_{2.95}$
LLDA-RFE	82.10 $^{(52)}_{1.38}$	81.30 $^{(60)}_{1.67}$	78.84 $^{(17)}_{3.39}$	**77.26** $^{(17)}_{7.19}$	**75.48** $^{(15)}_{4.78}$
DEODA	82.49 $^{(57)}_{1.45}$	81.30 $^{(60)}_{1.67}$	78.96 $^{(58)}_{4.36}$	74.26 $^{(60)}_{2.99}$	69.28 $^{(60)}_{3.90}$
All 60 features	81.91 $_{0.68}$	81.30 $_{1.67}$	78.72 $_{3.80}$	74.26 $_{2.99}$	69.28 $_{3.90}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	84.82 $^{(38)}_{3.19}$	81.64 $^{(54)}_{3.28}$	79.33 $^{(30)}_{4.83}$	76.66 $^{(35)}_{6.69}$	72.48 $^{(26)}_{2.26}$
SVD-Entropy	**85.10** $^{(51)}_{2.21}$	82.11 $^{(51)}_{3.28}$	80.45 $^{(57)}_{3.96}$	76.99 $^{(46)}_{2.90}$	74.50 $^{(49)}_{2.95}$
LLDA-RFE	82.62 $^{(54)}_{0.74}$	81.30 $^{(60)}_{1.67}$	79.05 $^{(59)}_{4.68}$	75.47 $^{(11)}_{10.19}$	74.55 $^{(9)}_{11.40}$
DEODA (ICs)	84.77 $^{(41)}_{2.98}$	**83.76** $^{(37)}_{11.39}$	**82.72** $^{(38)}_{2.00}$	**80.16** $^{(32)}_{3.14}$	**78.99** $^{(31)}_{3.44}$
All 60 features	81.91 $_{0.68}$	81.30 $_{1.67}$	78.72 $_{3.80}$	74.26 $_{2.99}$	69.28 $_{3.90}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

B) Sonar Data Set consists of two classes of 208 data with 60 features. In Tables 5.2 (A) and 5.2 (B), the accuracy of SVD-Entropy was higher than those of other methods for $k = 1, 3, 5$ using data space and $k = 1$ using transformed space. The degree of accuracy of LLDA-RFE was higher than those of other methods for $k = 7, 9$. DEODA performed better than those of the other methods for $k = 3, 5, 7, 9$ using the transformed space. Note that this data set was previously experimented in [5] by using supervised feature selection and classifying by multilayer perceptron (MLP) neural networks. They obtained the accuracy ranging in between 80-87% with 15 features.

Table 5.3: Average classification rate of selected feature subset using information from different spaces of the Parkinsons data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	83.90 $^{(6)}_{2.23}$	83.97 $^{(7)}_{2.63}$	84.33 $^{(7)}_{2.47}$	82.95 $^{(4)}_{2.11}$	81.78 $^{(5)}_{1.82}$
SVD-Entropy	83.90 $^{(22)}_{2.12}$	83.97 $^{(22)}_{2.63}$	84.43 $^{(21)}_{1.88}$	83.25 $^{(21)}_{1.41}$	81.72 $^{(22)}_{2.22}$
LLDA-RFE	83.90 $^{(19)}_{1.94}$	83.97 $^{(19)}_{2.63}$	84.54 $^{(2)}_{2.44}$	83.88 $^{(2)}_{1.88}$	83.62 $^{(2)}_{3.39}$
DEODA	83.90 $^{(22)}_{2.12}$	**84.38** $^{(13)}_{1.48}$	**85.31** $^{(8)}_{0.59}$	**84.13** $^{(8)}_{3.45}$	**83.63** $^{(8)}_{2.81}$
All 22 features	83.90 $_{2.12}$	83.97 $_{2.63}$	84.33 $_{2.47}$	82.69 $_{1.10}$	81.72 $_{2.22}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	85.16 $^{(7)}_{4.50}$	85.32 $^{(10)}_{2.34}$	85.67 $^{(12)}_{3.21}$	83.43 $^{(10)}_{2.52}$	82.08 $^{(17)}_{2.86}$
SVD-Entropy	83.90 $^{(22)}_{2.12}$	83.97 $^{(22)}_{2.63}$	84.43 $^{(21)}_{1.88}$	83.25 $^{(21)}_{1.41}$	81.72 $^{(22)}_{2.22}$
LLDA-RFE	84.11 $^{(20)}_{2.57}$	83.97 $^{(22)}_{2.63}$	84.43 $^{(18)}_{2.29}$	82.70 $^{(13)}_{3.14}$	82.44 $^{(4)}_{1.35}$
DEODA (ICs)	**89.01** $^{(15)}_{1.12}$	**89.15** $^{(15)}_{1.74}$	**88.54** $^{(20)}_{1.24}$	**89.05** $^{(20)}_{1.47}$	**88.43** $^{(20)}_{1.41}$
All 22 features	83.90 $_{2.12}$	83.97 $_{2.63}$	84.33 $_{2.47}$	82.69 $_{1.10}$	81.72 $_{2.22}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

C) Parkinsons Data Set consists of two classes of 195 data with 22 features. In Tables 5.3 (A) and 5.3(B), the accuracy of DEODA algorithm on data space was comparable or slightly superior to those of the other methods. In addition, when using the information from transformed space, the accuracy of the selected feature subset from DEODA was also superior to the others' regardless of the numbers of nearest neighbours.

Table 5.4: Average classification rate of selected feature subset using information from different spaces of the Ionosphere data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	88.37 $^{(18)}_{1.17}$	87.40 $^{(17)}_{0.38}$	85.72 $^{(17)}_{0.60}$	84.93 $^{(17)}_{0.77}$	84.30 $^{(20)}_{0.37}$
SVD-Entropy	86.75 $^{(28)}_{0.78}$	85.12 $^{(30)}_{0.32}$	84.47 $^{(30)}_{0.09}$	83.64 $^{(30)}_{0.18}$	83.22 $^{(33)}_{0.17}$
LLDA-RFE	87.32 $^{(12)}_{1.86}$	86.27 $^{(10)}_{2.40}$	84.96 $^{(6)}_{2.54}$	83.99 $^{(6)}_{1.67}$	83.56 $^{(32)}_{0.43}$
DEODA	**91.11** $^{(7)}_{0.56}$	**90.85** $^{(6)}_{0.87}$	**91.20** $^{(5)}_{1.58}$	**90.66** $^{(5)}_{0.50}$	**89.83** $^{(6)}_{1.50}$
All 33 features	86.32 $_{0.43}$	84.52 $_{0.49}$	84.16 $_{0.09}$	83.33 $_{0.17}$	83.22 $_{0.17}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	86.61 $^{(17)}_{1.71}$	86.95 $^{(13)}_{1.15}$	**86.55** $^{(13)}_{0.41}$	**85.30** $^{(13)}_{1.03}$	**84.98** $^{(14)}_{0.38}$
SVD-Entropy	86.75 $^{(28)}_{0.78}$	85.12 $^{(30)}_{0.32}$	84.47 $^{(30)}_{0.09}$	83.64 $^{(30)}_{0.18}$	83.22 $^{(33)}_{0.17}$
LLDA-RFE	86.72 $^{(22)}_{0.98}$	85.90 $^{(8)}_{3.31}$	84.85 $^{(5)}_{2.52}$	83.73 $^{(6)}_{2.01}$	83.47 $^{(32)}_{0.56}$
DEODA (ICs)	**88.14** $^{(18)}_{2.04}$	**87.26** $^{(9)}_{1.17}$	86.30 $^{(7)}_{0.82}$	85.24 $^{(8)}_{0.88}$	84.10 $^{(11)}_{1.15}$
All 33 features	86.32 $_{0.43}$	84.52 $_{0.49}$	84.16 $_{0.09}$	83.33 $_{0.17}$	83.22 $_{0.17}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

D) Ionosphere Data Set consists of two classes of 351 instances with 33 features. In Tables 5.4 (A) and 5.4(B), the accuracy of the DEODA algorithm was superior to the others' regardless of the numbers of nearest neighbours in the data space. In the transformed space, the accuracy of the selected feature subset from Laplacian score was higher than those of other methods for $k = 5, 7, 9$ but the selected feature subset from DEODA algorithm achieved the highest accuracy for $k = 1$.

Table 5.5: Average classification rate of selected feature subset using information from different spaces of the Soybean data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	**100.0** $^{(13)}_{0.00}$	99.80 $^{(13)}_{0.45}$	**100.0** $^{(13)}_{0.00}$	98.49 $^{(13)}_{3.07}$	95.98 $^{(14)}_{0.45}$
SVD-Entropy	**100.0** $^{(20)}_{0.00}$	**100.0** $^{(20)}_{0.00}$	**100.0** $^{(20)}_{0.00}$	98.93 $^{(20)}_{2.26}$	95.98 $^{(20)}_{0.45}$
LLDA-RFE	98.09 $^{(21)}_{0.45}$	99.16 $^{(21)}_{1.21}$	98.07 $^{(18)}_{2.46}$	96.60 $^{(21)}_{2.21}$	91.89 $^{(21)}_{7.85}$
DEODA	**100.0** $^{(15)}_{0.00}$	**100.0** $^{(14)}_{0.00}$	**100.0** $^{(15)}_{0.00}$	**99.80** $^{(16)}_{0.45}$	**99.78** $^{(15)}_{0.45}$
All 35 features	98.09 $_{0.45}$	99.16 $_{1.21}$	97.44 $_{0.80}$	96.40 $_{3.07}$	91.47 $_{6.04}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	99.16 $^{(17)}_{1.21}$	99.40 $^{(17)}_{1.06}$	99.78 $^{(14)}_{0.45}$	97.44 $^{(20)}_{2.82}$	94.89 $^{(14)}_{7.24}$
SVD-Entropy	**100.0** $^{(20)}_{0.00}$	**100.0** $^{(20)}_{0.00}$	**100.0** $^{(20)}_{0.00}$	98.93 $^{(20)}_{2.26}$	95.98 $^{(20)}_{0.45}$
LLDA-RFE	98.09 $^{(21)}_{0.45}$	99.58 $^{(20)}_{0.80}$	98.38 $^{(16)}_{6.84}$	97.64 $^{(15)}_{2.46}$	95.31 $^{(15)}_{3.82}$
DEODA (ICs)	**100.0** $^{(9)}_{0.00}$	**100.0** $^{(12)}_{0.00}$	**100.0** $^{(12)}_{0.00}$	**100.0** $^{(9)}_{0.00}$	**100.0** $^{(12)}_{0.00}$
All 35 features	98.09 $_{0.45}$	99.16 $_{1.21}$	97.44 $_{0.80}$	96.40 $_{3.07}$	91.47 $_{6.04}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

E) Soybean Data Set consists of four classes of 47 instances with 35 features. In Tables 5.5 (A) and 5.5 (B), the accuracy of the DEODA algorithm was superior to the others' regardless of the numbers of nearest neighbours when being applied directly on the individual features in the original data space and the transformed space.

Table 5.6: Average classification rate of selected feature subset using information from different spaces of the SRBCT data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	89.38 $^{(2256)}_{3.39}$	83.40 $^{(2219)}_{5.71}$	79.88 $^{(1928)}_{5.63}$	78.32 $^{(1600)}_{5.71}$	76.92 $^{(1433)}_{8.54}$
SVD-Entropy	93.42 $^{(1188)}_{2.13}$	88.75 $^{(1260)}_{7.84}$	86.79 $^{(1224)}_{6.19}$	84.59 $^{(1415)}_{10.67}$	81.76 $^{(1185)}_{18.62}$
LLDA-RFE	89.70 $^{(1847)}_{4.06}$	84.23 $^{(1146)}_{9.52}$	80.78 $^{(1489)}_{3.05}$	77.67 $^{(1139)}_{11.76}$	76.53 $^{(988)}_{23.54}$
DEODA	**100.0** $^{(74)}_{0.00}$	**99.71** $^{(73)}_{0.45}$	**98.89** $^{(83)}_{1.71}$	**98.44** $^{(68)}_{2.80}$	**98.25** $^{(83)}_{2.49}$
All 2308 features	88.79 $_{2.49}$	82.62 $_{8.65}$	78.45 $_{4.62}$	75.32 $_{13.02}$	73.42 $_{10.67}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	88.79 $^{(2295)}_{2.49}$	83.22 $^{(2180)}_{7.42}$	78.77 $^{(2289)}_{9.55}$	75.63 $^{(2305)}_{12.91}$	73.42 $^{(2308)}_{10.67}$
SVD-Entropy	**93.42** $^{(1188)}_{2.13}$	**88.75** $^{(1260)}_{7.84}$	**86.79** $^{(1224)}_{6.19}$	**84.59** $^{(1415)}_{10.67}$	**81.76** $^{(1185)}_{18.62}$
LLDA-RFE	89.87 $^{(1461)}_{7.95}$	83.62 $^{(1417)}_{11.23}$	80.33 $^{(913)}_{14.59}$	78.33 $^{(210)}_{15.71}$	76.86 $^{(146)}_{20.86}$
DEODA (ICs)	89.55 $^{(2291)}_{2.91}$	83.92 $^{(2270)}_{6.97}$	78.46 $^{(2279)}_{5.18}$	77.26 $^{(1905)}_{11.23}$	76.57 $^{(450)}_{15.15}$
All 2308 features	88.79 $_{2.49}$	82.62 $_{8.65}$	78.45 $_{4.62}$	75.32 $_{13.02}$	73.42 $_{10.67}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

F) SRBCT Data Set [37] consists of four distinct diagnostic categories of small and round blue tumor cells. It is gene-expression data from cDNA micro-arrays containing 2,308 genes and 63 samples. All genes were normalized to zero mean and unit variance. The degree of accuracy of SVD-Entropy was the highest, i.e. 93.42%, with 1,188 genes in the transformed space as shown in Table 5.6 (B). However, the degree of accuracy of the DEODA algorithm was superior to the others's regardless of the numbers of nearest neighbours in the data space as shown in Table 5.6 (A). Moreover, the accuracy of selected feature subset from DEODA using one nearest neighbour reached 100% with a subset of 74 features. Obviously, this feature subset was smaller than the subset of size 94 features found in [37] using a supervised feature selection.

Table 5.7: Average classification rate of selected feature subset using information from different spaces of the ALL-AML data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	96.14 $^{(1962)}_{3.46}$	94.85 $^{(1612)}_{7.69}$	90.29 $^{(551)}_{9.31}$	87.95 $^{(466)}_{9.54}$	**87.62** $^{(383)}_{7.77}$
SVD-Entropy	87.99 $^{(5981)}_{3.39}$	91.13 $^{(6906)}_{4.92}$	86.16 $^{(5578)}_{6.23}$	81.61 $^{(5717)}_{1.54}$	80.86 $^{(4971)}_{1.62}$
LLDA-RFE	93.74 $^{(902)}_{9.54}$	93.67 $^{(1923)}_{12.62}$	89.78 $^{(480)}_{9.93}$	89.13 $^{(420)}_{6.85}$	86.54 $^{(272)}_{13.00}$
DEODA	**97.38** $^{(166)}_{4.62}$	**98.93** $^{(463)}_{3.39}$	**95.29** $^{(444)}_{5.85}$	**91.89** $^{(380)}_{6.85}$	86.73 $^{(501)}_{3.77}$
All 7129 features	86.71 $_{3.77}$	90.88 $_{6.54}$	85.08 $_{3.15}$	80.35 $_{6.54}$	79.07 $_{3.08}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	88.28 $^{(6030)}_{1.92}$	90.88 $^{(6956)}_{6.54}$	85.87 $^{(6998)}_{4.92}$	81.42 $^{(2991)}_{3.77}$	**80.92** $^{(4311)}_{3.15}$
SVD-Entropy	87.99 $^{(5981)}_{3.39}$	91.13 $^{(6906)}_{4.92}$	**86.16** $^{(5578)}_{6.23}$	81.61 $^{(5717)}_{1.54}$	80.86 $^{(4971)}_{1.62}$
LLDA-RFE	**90.63** $^{(1487)}_{8.00}$	**91.67** $^{(1177)}_{15.08}$	85.93 $^{(786)}_{6.46}$	**83.38** $^{(328)}_{12.39}$	80.58 $^{(610)}_{8.00}$
DEODA (ICs)	88.24 $^{(5681)}_{3.46}$	90.88 $^{(7077)}_{6.54}$	85.08 $^{(6933)}_{3.15}$	80.35 $^{(6785)}_{6.54}$	79.07 $^{(7038)}_{3.08}$
All 7129 features	86.71 $_{3.77}$	90.88 $_{6.54}$	85.08 $_{3.15}$	80.35 $_{6.54}$	79.07 $_{3.08}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

G) ALL-AML Data Set [38] consists of two classes of 38 leukemia cell samples and 7,129 genes. All genes were normalized to zero mean and unit variance. The degree of accuracy of LLDA-RFE was the highest, i.e. 91.67%, with 1,177 genes in the transformed space as shown in Table 5.7 (B). However, the best subset was the subset of 463 genes indicated by DEODA algorithm with accuracy of 98.93% using 3-nearest neighbour algorithm in the data space as shown in Table 5.7 (A).

Table 5.8: Average classification rate of selected feature subset using information from different spaces of the MLL data set

(A) original data space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	88.20 $^{(1195)}_{2.59}$	86.61 $^{(3821)}_{2.66}$	**87.09** $^{(9096)}_{3.88}$	84.32 $^{(9842)}_{8.77}$	81.14 $^{(545)}_{6.02}$
SVD-Entropy	83.45 $^{(11866)}_{3.94}$	81.89 $^{(12321)}_{3.45}$	82.59 $^{(12481)}_{1.48}$	81.48 $^{(12466)}_{9.54}$	77.32 $^{(12324)}_{9.19}$
LLDA-RFE	89.41 $^{(2029)}_{1.71}$	86.80 $^{(1040)}_{2.25}$	86.38 $^{(1332)}_{2.16}$	84.67 $^{(7370)}_{5.59}$	82.86 $^{(518)}_{6.02}$
DEODA	**91.58** $^{(1072)}_{1.63}$	**88.33** $^{(1354)}_{0.94}$	86.42 $^{(7590)}_{2.59}$	**86.42** $^{(2641)}_{7.31}$	**85.53** $^{(1878)}_{1.63}$
All 12582 features	83.27 $_{3.54}$	81.89 $_{3.45}$	82.59 $_{1.48}$	81.30 $_{9.09}$	76.79 $_{11.06}$

(B) transformed space **Accuracy of nearest neighbour classifier**

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Laplacian Score	83.45 $^{(10900)}_{3.09}$	82.24 $^{(12140)}_{4.48}$	82.59 $^{(12495)}_{1.48}$	81.47 $^{(12463)}_{9.11}$	77.33 $^{(12429)}_{8.34}$
SVD-Entropy	83.45 $^{(11866)}_{3.94}$	81.89 $^{(12321)}_{3.45}$	82.59 $^{(12481)}_{1.48}$	81.48 $^{(12466)}_{9.54}$	77.32 $^{(12324)}_{9.19}$
LLDA-RFE	**84.15** $^{(7440)}_{4.72}$	**83.59** $^{(4019)}_{4.31}$	**84.50** $^{(7647)}_{2.49}$	**82.88** $^{(8364)}_{9.88}$	**79.79** $^{(6801)}_{9.97}$
DEODA (ICs)	83.64 $^{(12406)}_{3.45}$	82.41 $^{(12368)}_{1.71}$	82.59 $^{(12451)}_{1.48}$	81.67 $^{(12140)}_{8.23}$	77.32 $^{(12362)}_{9.19}$
All 12582 features	83.27 $_{3.54}$	81.89 $_{3.45}$	82.59 $_{1.48}$	81.30 $_{9.09}$	76.79 $_{11.06}$

* the superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

H) MLL Data Set [39] consists of three kinds of leukemia samples with 12,582 genes and 57 samples. All genes were normalized to zero mean and unit variance. The data set is available at [40]. The degree of accuracy of SVD-Entropy was the highest, i.e. 84.50%, with 7,647 genes in the transformed space as shown in Table 5.8 (B). However, the best subset wa s the subset of 1,072 genes indicated by DEODA algorithm with accuracy of 91.58% using 1-nearest neighbour algorithm in the data space as shown in Table 5.8 (A). Note that this data set was also used in [6]. They filtered the genes before applying feature selection and the degree of accuracy ranged in between 94-96% using the Nearest Mean Classifier (NMC).

To compare the performance of every algorithm, an average performance evaluation is introduced. The accuracy using all original features are used as a base line performance and the performance of every algorithm is computed with respect to the base line. The average

performance evaluation of algorithms on j^{th} data set can be computed by

$$Q_j = \frac{1}{n_{ex}} \sum\nolimits_k Q_{j,k}^{(m^*)} / Q_{j,k}^{(All)} \times 100\% \tag{5.1}$$

where n_{ex} is the total number of experiments using different k, in this case $n_{ex} = 5$. $Q_{j,k}^{(m^*)}$ and $Q_{j,k}^{(All)}$ are the average accuracy of k nearest neighbours of selected subset size m^* and all original features of the j^{th} data set, respectively. These can be taken from Tables 5.1 - 5.8. Then, the overall average performances over the base line can be computed by averaging the performances over the base line as follows

$$Q = \frac{1}{n_{data}} \sum\nolimits_j Q_j \tag{5.2}$$

where n_{ex} is the total number of data set.

The average performance of all algorithms when using the information from original space over the base line were shown in Figure 5.7. For WDBC data set, all algorithms achieve the mean performance comparable with the other methods. LLDA-RFE and SVD-Entropy have the average performance over the base line higher than that using DEODA for Sonar data set. However, the average performances over the base line of DEODA are superior for Parkinsons, Ionosphere and Soybean data sets. Figure 5.8 shows the average performances of all algorithms in the transformed space over the base line. Observe that the average performances over the base line of DEODA were superior for all data sets. The average performance of all algorithms when using the information from original space over the base line on the large feature data set were shown in Figure 5.10. The average performances over the base line of DEODA were superior for all data sets.

The overall average performances over the base line (performances for short) of all algorithms on the UCI data set were shown in Figure 5.9. The performance of DEODA algorithm was higher than those of the other algorithms when using the information from both data space and transformed space. The performances of all algorithms on the large feature data set were shown in Figure 5.12. The performances of all algorithms on data space were higher than those using the information on the transformed space. On the transformed space, the performances of SVD-Entropy was higher than those of the other algorithms. However, the performances of DEODA algorithm was higher than those of the other algorithms when using the information from the data space and much higher than those of using the transformed space. These results indicated that the prediction accuracy can be improved by using feature selection algorithms.

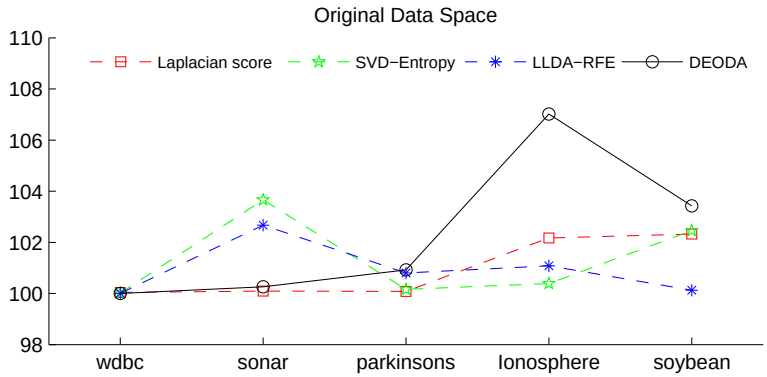

Figure 5.7: Average performance of every algorithm over the base line when using the information from the given data space of the UCI data sets.

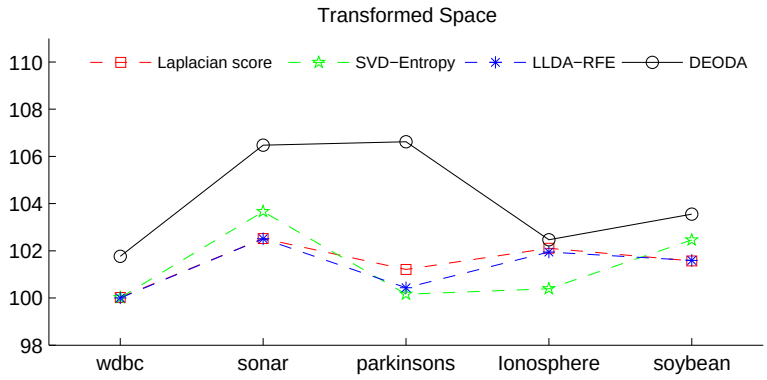

Figure 5.8: Average performance of every algorithm over the base line when using the information from the transformed space of the UCI data sets.

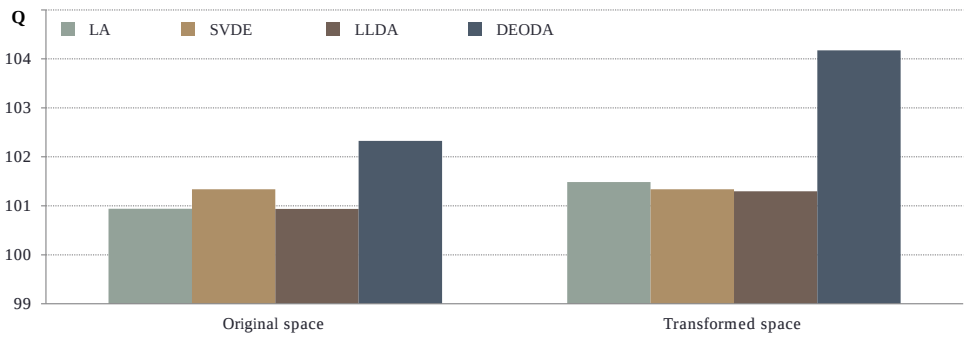

Figure 5.9: The overall average performance over the based line of every algorithms of the UCI data sets.

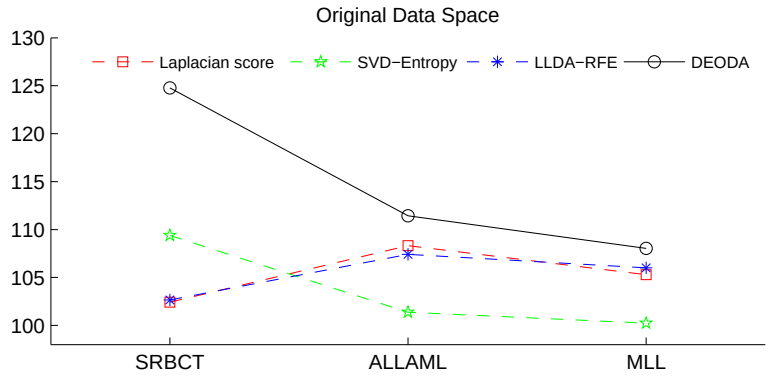

Figure 5.10: Average performance of every algorithm over the base line when using the information from the given data space of the micro-array data sets.

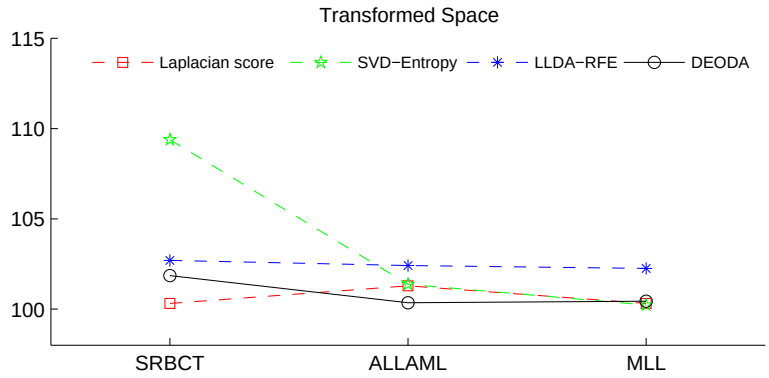

Figure 5.11: Average performance of every algorithm over the base line when using the information from the transformed space of the micro-array data sets.

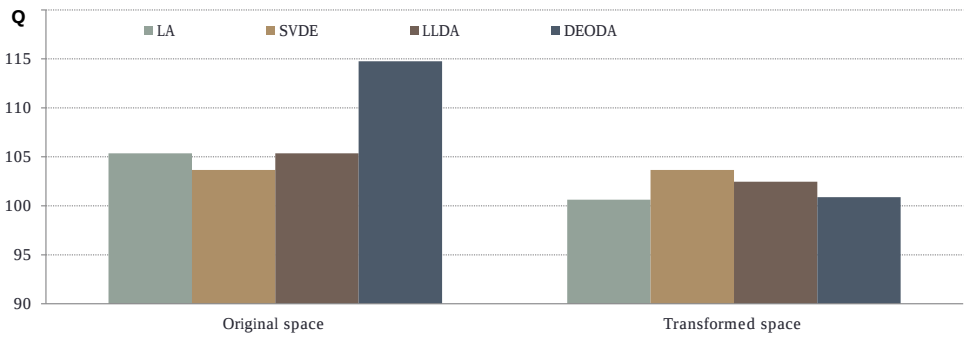

Figure 5.12: The overall average performance over the based line of every algorithms of the micro-array data sets.

5.4 Performance Measurement on Clustering

In this section, we applied feature selection and clustering on the gene micro-array data sets because the analysis in the previous studies was usually based on cluster analysis. Because, the clustering results reflect true structure in the data [39] which is constrained by clustering algorithm. In addition, the data distribution of the best gene subset was rather clearly separated as reported in [37–39]. The hierarchical clustering algorithm based on correlation distance and average link strategy as in [37] was used. The clustering experiments were conducted only on the data space. Since, the results as shown in Figure 5.12 indicated that, for these data sets, it was appropriate to apply the algorithms on the original data space.

Let us introduce an index for selecting the proper genes subset from the candidate subsets. The Calinski-Harabasz Index (CH-index) [41] is one of cluster separation measures. The maximum value of the CH-index indicates that the clusters are mostly separated among each other and the data in cluster are highly dense distribution. The CH-index is an unsupervised index and can be computed as follows

$$CH = \frac{tr(\mathbf{S}_B)}{tr(\mathbf{S}_W)} \cdot \frac{G-1}{n-G} \tag{5.3}$$

where $tr(\mathbf{S}_W) = \sum_{g=1}^{G} \sum_{i=1}^{n^{(g)}} ||\mathbf{x}_i^{(g)} - \bar{\mathbf{x}}^{(g)}||^2$, and $tr(\mathbf{S}_B) = \sum_{g=1}^{G} ||\bar{\mathbf{x}}^{(g)} - \bar{\mathbf{x}}||^2$, G is the number of clusters, $\mathbf{x}_i^{(g)}$ denotes the i^{th} instance which presents in the g^{th} cluster, $\bar{\mathbf{x}}^{(g)}$ is the mean vector of the g^{th} cluster, $\bar{\mathbf{x}}$ is the mean vector of the the given data, $n^{(g)}$ is the number of memberships of the g^{th} cluster. However, any cluster with a few instances being viewed as an outlier instances, can maximize the (5.3). This issue can be reduced by multiplying (5.3) with the entropy of number of clusters members. Therefore, the modified CH-index is as follows

$$mCH = (-\sum_{g=1}^{G} n^{'(g)} \log n^{'(g)})CH \tag{5.4}$$

where $n^{'(g)} = n^{(g)} / \sum_g n^{(g)}$ is the normalized membership number of the g^{th} cluster.

The experiments were conducted in two different ways as follows.

5.4.1 Genes Selection for Class Discovery

This evaluation was conducted based on the assumption that the given data set consisted of irrelevant genes which can negatively affect the cluster analysis. Therefore, it also was expected

that the clustering result of selected genes subset should be better than the result based on all given genes. The experiment consisted of the following steps.

1. The genes were rearranged by each feature selection algorithm.
2. The hierarchical clustering was applied to all m candidate subsets selected from the ranked genes. The number of clusters was set equal to the number of classes.
3. The mCH-index was applied for selecting the appropriated genes subset of size m^*.
4. The cluster labels of the selected genes subset were compared with known class labels [37–39].

The clustering results were shown in Table 5.9. The second column was the results of gene subsets selected by the modified CH-index. The results of the fixed subset size, which was equal to the subset size of the best subset selected by the modified CH-index, were shown in the third column. The maximum results were also shown in the last column.

The clustering results of the selected feature subset using the DEODA algorithm were much higher than that using other alternative methods as well as all original genes as shown in Table 5.9. There were results indicated that a feature subset existed such that the clustering results were exactly the same with the given class label. These were the feature subsets indicated by the DEODA algorithm on SRBCT and ALL-AML data set. Note that our algorithm is fully unsupervised. Moreover, the maximum accuracy of clustering results of all algorithms were also higher than the original features. This supported the assumption that the data set consists of some irrelevant features that negatively affect the cluster analysis. Figure5.13 are the performance of clustering results of all feature subset sizes. The clustering accuracy of feature subsets indicated by the DEODA algorithm were higher than those using the other algorithms. Moreover, the relevant genes were ranked in the top order since the clustering accuracy of small feature subset sizes were higher than the large subset sizes as well as all original genes.

5.4.2 Genes Clustering

This experiment was conducted for finding the representative sample subset for gene clustering. In this experiment, each gene can be viewed as an data and each sample can be viewed as a feature. Note that since the alternative methods used information from the given data, it was generally expected to perform better than our algorithm for finding the representative feature subset. The experiment is as follows.

Table 5.9: Clustering performance of the micro-array data set using information from the original data space. The superscript next to the performance denote the feature subset size.

(A) SRBCT Clustering accuracy

Algorithm	Selected subset	fixed subset size	Maximum
Laplacian Score	$55.56^{(4)}$	$52.38^{(10)}$	$57.14^{(1536)}$
SVD-Entropy	$53.97^{(2308)}$	$36.51^{(10)}$	$95.24^{(767)}$
LLDA-RFE	$41.27^{(21)}$	$42.86^{(10)}$	$60.32^{(4)}$
DEODA	$87.3^{(10)}$	$87.3^{(10)}$	$100^{(49)}$
All original 2308 genes 53.97			

(B) ALL-AML Clustering accuracy

Algorithm	Selected subset	fixed subset size	Maximum
Laplacian Score	$84.21^{(20)}$	$68.42^{(36)}$	$89.47^{(41)}$
SVD-Entropy	$65.79^{(2)}$	$68.42^{(36)}$	$100^{(2831)}$
LLDA-RFE	$71.05^{(3)}$	$73.68^{(36)}$	$76.32^{(2)}$
DEODA	$94.74^{(36)}$	$94.74^{(36)}$	$100^{(65)}$
All original 7129 genes 76.32			

(C) MLL Clustering accuracy

Algorithm	Selected subset	fixed subset size	Maximum
Laplacian Score	$68.42^{(4)}$	$71.93^{(394)}$	$78.95^{(5219)}$
SVD-Entropy	$43.86^{(8)}$	$40.35^{(394)}$	$75.44^{(11195)}$
LLDA-RFE	$80.7^{(46)}$	$73.68^{(394)}$	$80.7^{(46)}$
DEODA	$91.23^{(394)}$	$91.23^{(394)}$	$98.25^{(100)}$
All original 12582 genes 73.68			

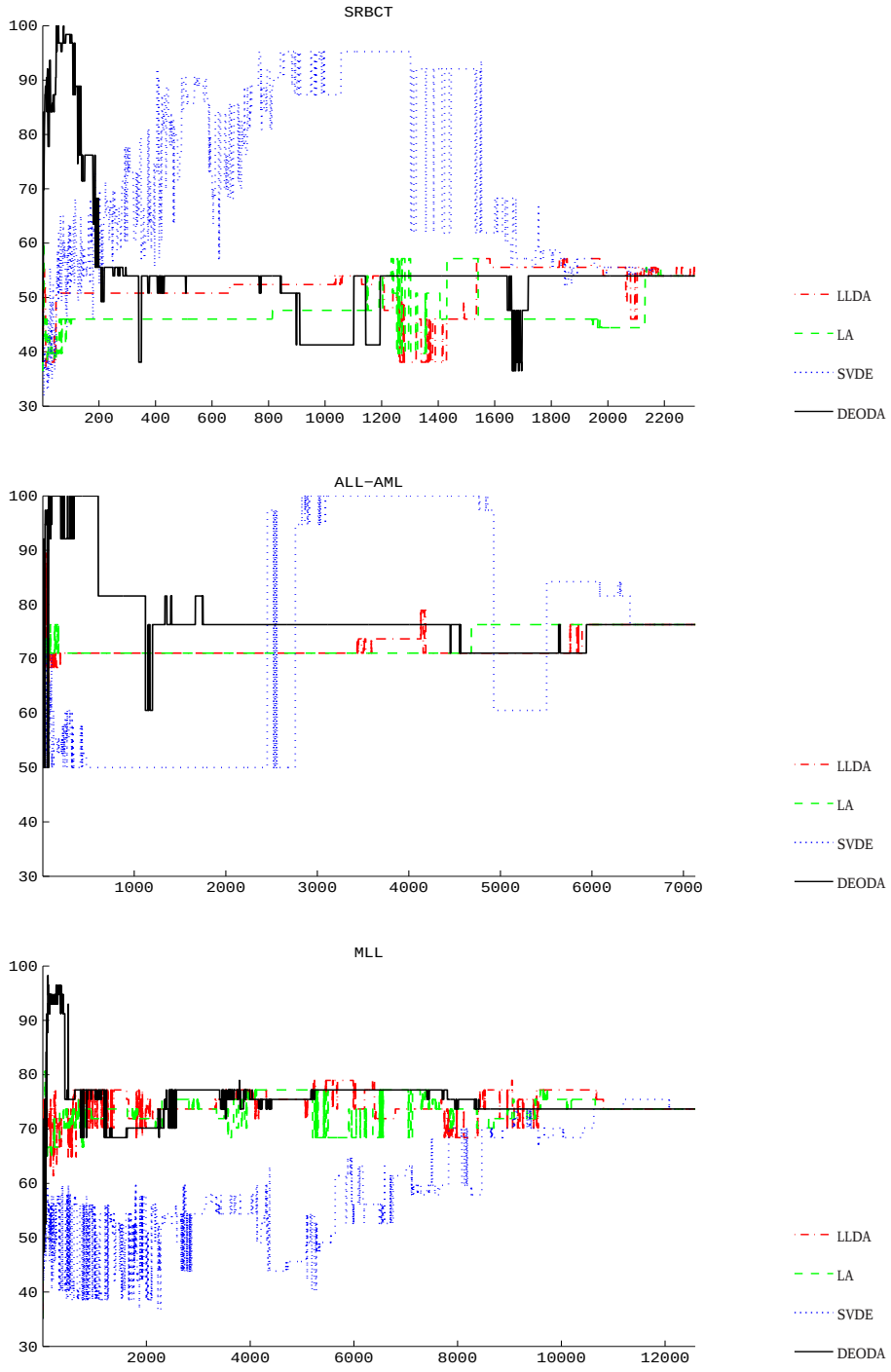

Figure 5.13: The performance of clustering results compared to the class label of all feature subsets from different algorithms. The vertical axis is the clustering performances while the horizontal axis is feature subset sizes.

1. The genes were clustered with the numbers of clusters equal to 2, 3, ..., 10, respectively.
2. The mCH-index was applied to select the appropriated number of clusters G^* and, then, the cluster labels were used as cluster targets.
3. The features were rearranged by each feature selection algorithm.
4. All candidate feature subsets (as in section VI) selected from the ranking were clustered as in step 1 with number of clusters equal to G^*.
5. The cluster labels of all candidate subsets were compared with the cluster targets in step 2.

Figure 5.14 showed the clustering accuracy of all feature subset sizes. We can observe the correctness of clustering results using DEODA algorithm were comparable with other methods which use some information from all original features.

5.4.3 Parameter Sensitivity

This experiment was conducted for comparing the performance of using different number of histogram bins N and different number of discrete Fourier transform samples B. We conducted experiments as in section VI-A with both UCI and Micro-array data set and, then computed the overall performance as (5.2). Firstly, B was restricted to 1024 while the numbers of histogram bins N was set to 8, 32, 64, 128, 256, and 512, respectively. The overall average performance were shown in Figure5.15 (a). We can observed that the number of bins should be set in the range of 32 to 256. Secondly, N was restricted to 32 while B was set to 32, 64, 128, 256, 512, 1024, 2048, and 4096, respectively. The overall average performance were shown in Figure5.15 (b). We can observed that the number of discrete Fourier transform samples should be set higher than 128 or four times the number of histogram bins.

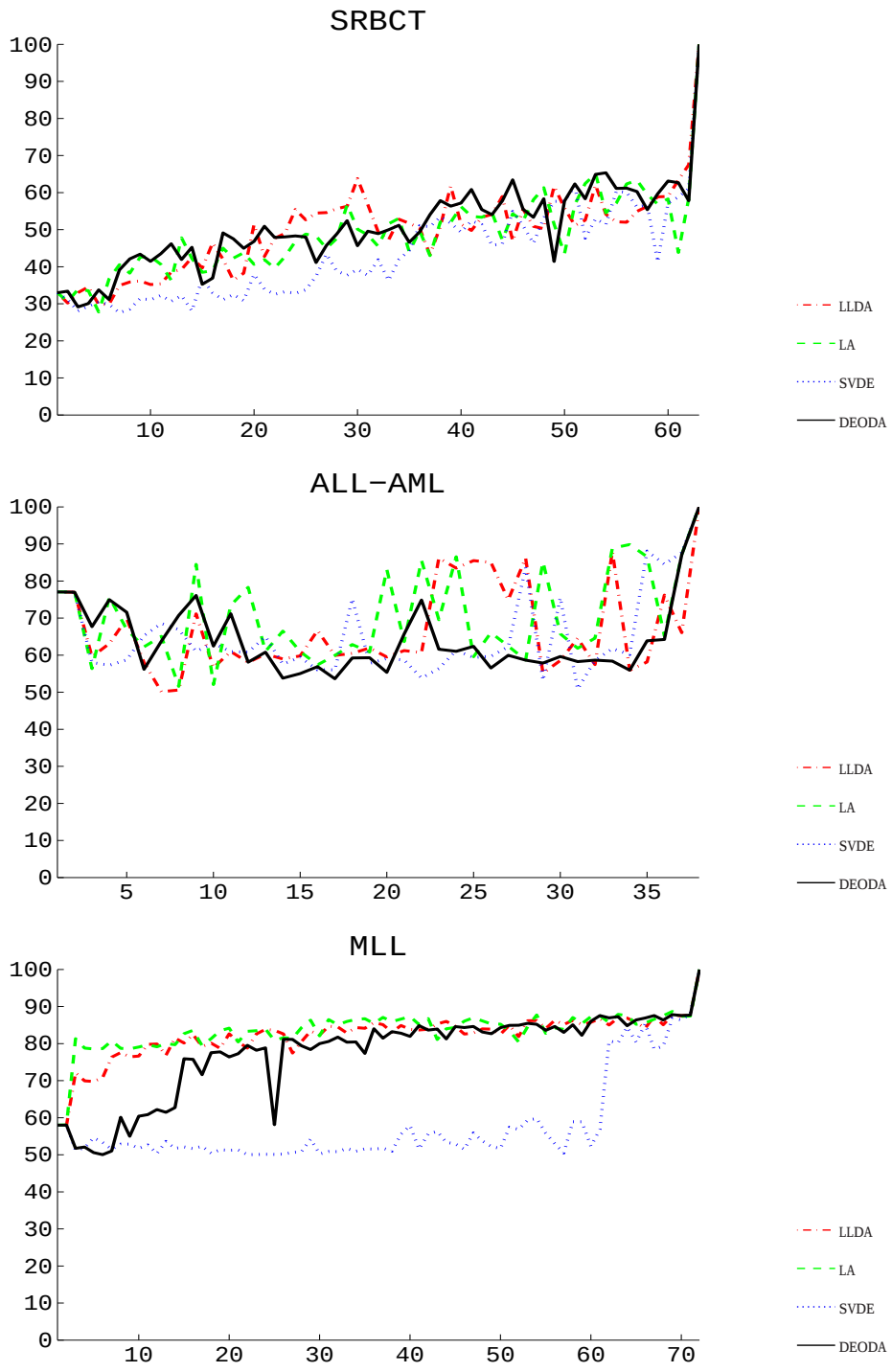

Figure 5.14: The performance of clustering results compared to the clustering result of all features from different algorithms. The vertical axis are the clustering performances while the horizontal axis are feature subset sizes.

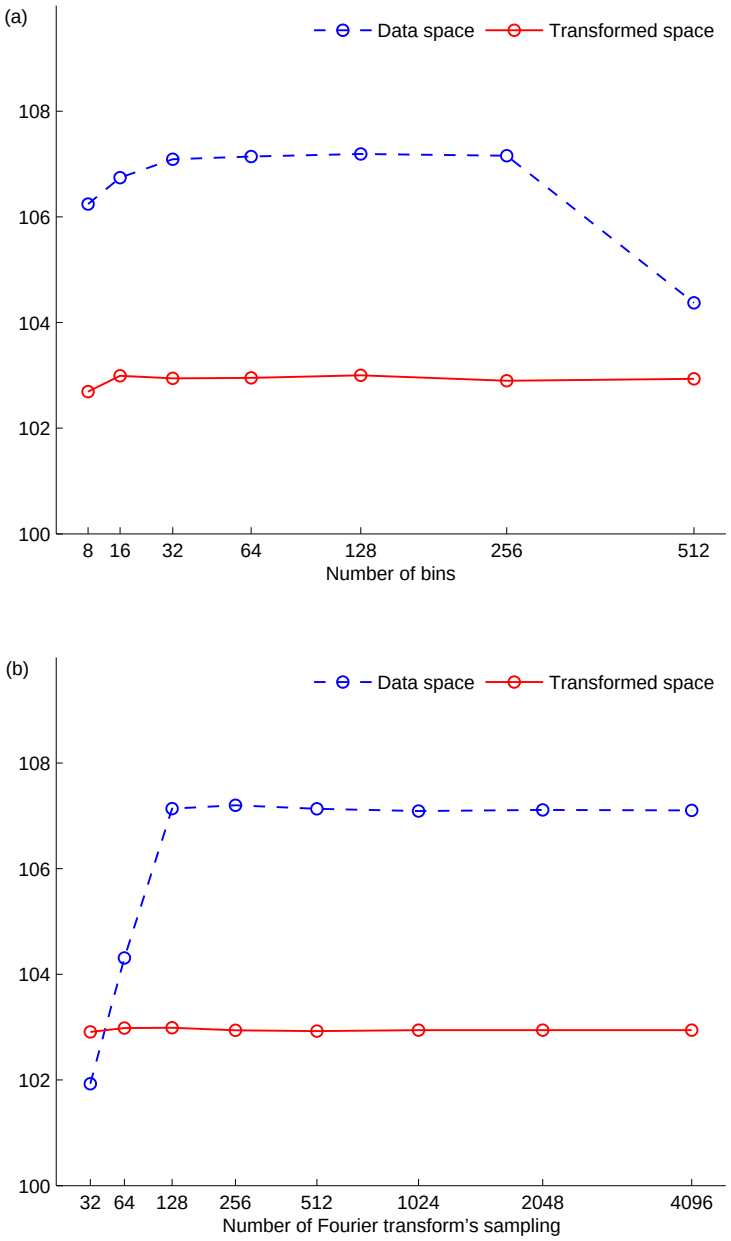

Figure 5.15: The overall average performance over the based line of the DEODA algorithm (a) using different number of histogram's bin, (b) using different number of discrete Fourier transform's sampling.

CHAPTER VI

DISCUSSION AND CONCLUSION

6.1 Discussion

Experimental results have shown the usefulness of the features selection algorithms. In most cases, feature subsets selected by each algorithm achieved a higher performance than the one using all of the original features. This indicated that the features of the given data were not always relevant because the selected features were a subset of the original features from the given data. Therefore, using all given features without removing irrelevant features can negatively affect the performance of the analysis, as shown in both the classification and clustering experiments.

According to the experimental results, the discrimination analysis via a diffraction pattern of a synthetic aperture is useful for unsupervised feature selection. In this dissertation, the proposed DEODA algorithm achieved a higher performance for both the classification and clustering than those of the other algorithms and also higher than using all of the original features. For Ionosphere and gene micro-array data set, the performance using the given data space was better than that of the transformed space. This indicated that the relevant discriminative distributions were found in data space, as illustrated in chapter V. But the transformation can negatively affect the discriminative capability of the original features. One possible reason is that the transformed space is not the optimum space required by the objective function of the transformation algorithm. Since the learning processes are needed to initiate the basis, which are randomly selected. The algorithm can convert to a local optimum that has more effect on higher dimension data, as shown in the micro-array results. Contrast functions are another possible reason. There are many ICA approaches with different contrast functions and different learning algorithms. This can have different effects on the discriminative capability of the original features. A fast fixed-point algorithm using Kurtosis as a contrast function was applied in this dissertation. Other approaches can be found in [43, 44] and some implemented codes can be found in [45].

Although the performance of selected features using DEODA algorithm was rather high, it still depends on a probability density estimation which is still an open problem for optimal

solution. In this dissertation, only a simple algorithm for density estimation was implemented. Another probability density estimation algorithm can be found in [32]. But when this algorithm [32] was used instead of the proposed density estimation, the performance of the feature subset selection was comparable to the proposed algorithm. Moreover, equation (4.3) can be viewed as a mapping function, $H = K(Y_\alpha)$. $K(\cdot)$ can be changed to any arbitrary mapping function, e.g. a sigmoidal function. These should be tested in order to achieve a higher performance. Also, an appropriate approach for reducing redundancy features can be combined in the score function (4.3) as in the supervised feature selection algorithm [5]. The experiments as in chapter 5.3 by using 3-NN on the UCI data sets were conducted. The results of combining the normalized mutual information criterion in the score function were shown in Table 6.1. However, it takes much more computation time while the number of features not much different and the performance of the selected feature subset were comparable to our algorithm.

Table 6.1 Average classification rate of algorithm with mutual information of the UCI data set

(A1) WDBC Accuracy of k-NN classifier (Data space)

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
DEODA	91.48 $^{(24)}_{0.37}$	92.89 $^{(30)}_{0.32}$	92.87 $^{(30)}_{0.26}$	93.24 $^{(30)}_{0.18}$	93.36 $^{(30)}_{0.09}$
DEODA with MI	**91.82** $^{(28)}_{0.54}$	92.89 $^{(30)}_{0.32}$	**92.91** $^{(29)}_{0.23}$	93.24 $^{(30)}_{0.18}$	93.36 $^{(30)}_{0.09}$

(B1) Sonar

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
DEODA	82.54 $^{(59)}_{1.54}$	81.30 $^{(60)}_{1.67}$	79.15 $^{(57)}_{4.12}$	**74.27** $^{(59)}_{3.86}$	69.28 $^{(60)}_{3.90}$
DEODA with MI	**82.58** $^{(57)}_{1.84}$	81.30 $^{(60)}_{1.67}$	**79.29** $^{(58)}_{5.01}$	74.26 $^{(60)}_{2.99}$	69.28 $^{(60)}_{3.90}$

(C1) Parkinsons

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
DEODA	83.90 $^{(22)}_{2.12}$	**84.38** $^{(10)}_{2.00}$	85.16 $^{(9)}_{1.87}$	83.93 $^{(7)}_{3.69}$	83.06 $^{(6)}_{3.53}$
DEODA with MI	83.90 $^{(22)}_{2.12}$	84.08 $^{(20)}_{3.24}$	**85.20** $^{(13)}_{1.25}$	**84.08** $^{(10)}_{4.00}$	**83.47** $^{(12)}_{2.98}$

(D1) Ionosphere

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
DEODA	**90.51** $^{(6)}_{1.26}$	90.71 $^{(6)}_{1.46}$	**90.74** $^{(5)}_{1.41}$	89.85 $^{(5)}_{1.21}$	89.26 $^{(4)}_{0.63}$
DEODA with MI	**90.51** $^{(6)}_{0.56}$	**90.77** $^{(6)}_{0.42}$	90.66 $^{(4)}_{0.79}$	**90.20** $^{(4)}_{0.58}$	**89.49** $^{(4)}_{0.66}$

(E1) Soybean

Algorithm	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
DEODA	**100.0** $^{(15)}_{0.00}$	**100.0** $^{(14)}_{0.00}$	**100.0** $^{(15)}_{0.00}$	**99.80** $^{(16)}_{0.45}$	**99.60** $^{(17)}_{0.80}$
DEODA with MI	99.80 $^{(15)}_{0.45}$	99.78 $^{(16)}_{0.45}$	**100.0** $^{(16)}_{0.00}$	**99.80** $^{(15)}_{0.45}$	99.38 $^{(14)}_{1.06}$

Note that, some results of DEODA algorithm may slightly different from the one shown in chapter 5.3 due to the random dividing. The superscript and subscript next to the performance denote the feature subset size and the standard deviation of the performance respectively

6.2 Conclusion

In this dissertation, a new discrimination analysis for unsupervised feature selection was proposed. The concepts of physical optics have been employed for discrimination evaluation of data distribution. The data distribution is assumed to be a synthetic aperture. Then, the far-field diffraction pattern was, then, observed by passing a plane of light waves through the synthetic aperture. The observation can be simulated using Fourier transform. Then, the magnitude of the Fourier transform of the synthetic aperture function is used as the intensity of diffraction pattern for discrimination analysis.

The discrimination of data distribution is observed by measuring the equality of the mode of magnitude in terms of the entropy of the first two highest modes of intensity. Moreover, these processes are analyzed with respect to a basis that can be rotated according to the distribution of the data. Then, the discriminative scores of all features are, then, computed using the information of the bases. If the bases are not standard bases, the discrimination analysis is performed on the transformed space produced by the bases.

The proposed algorithm can also be used as a univariate technique by restricting the bases equal to standard bases. The proposed feature evaluation algorithm evaluates the feature of the given data on original feature space in O(mn) when standard bases are used. Furthermore, the discrimination analysis does not depend on any scaling parameters because the distribution of the data does not change when the data are scaled. Experimental results on several real world data sets have demonstrated the effectiveness of the proposed method.

REFERENCES

[1] Torkkola, K. Feature Extraction by Non-Parametric Mutual Information Maximization. Machine Learning Research 3 (2003): 1415–1438.

[2] Ii, K. E. H., Erdogmus, D., Torkkola, K., and Principe, J. C. Feature Extraction Using Information-Theoretic Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 28 , 9 (2006): 1385–1392.

[3] Battiti, R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transactions on Neural Networks 5 , 4 (1994): 537–550.

[4] Yu, L. and Liu, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Machine Learning Research 5 (2004): 1205–1224.

[5] Estevez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. Normalized Mutual Information Feature Selection. IEEE Transactions on Neural Networks 20 , 2 (2009): 189–201.

[6] Niijima, S. and Okuno, Y. Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 6 , 4 (2009): 605–614.

[7] Bellman, R. E. Adaptive Control Processes - A guided tour. U.S.A. : Princeton University Press, 1961.

[8] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U., When Is ”Nearest Neighbor” Meaningful?”. Int. Conf. on Database Theory, Canada, (1999): 217–235.

[9] Saeys, Y., Inza, I., and Larranaga, P. A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics 23 , 19.

[10] Kohavi, R. and John, G. H. Wrappers for Feature Subset Selection. Artificial Intelligence 97 , 1–2 (1997): 273–324.

[11] Li, Y., Dong, M., and Hua, J. Simultaneous Localized Feature Selection and Model Detection for Gaussian Mixtures. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31 , 5 (2009): 953–960.

[12] Law, M. H., Figueiredo, M. A., and Jain, A. K. Simultaneous Feature Selection and Clustering Using Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 26 , 9 (2004): 1154–1166.

[13] Zeng, H. and Cheung, Y.-M. A new Feature Selection Method for Gaussian Mixture Clustering. Pattern Recognition 42 , 2 (2009): 243–250.

[14] Roy, V., Assaf, G., Michal, L., and David, H. Novel Unsupervised Feature Filtering of Biological Data. Bioinformatics 22 , 14 (2006): e507–e513.

[15] He, X., Cai, D., and Niyogi, P., Laplacian Score for Feature Selection. in *Advances in Neural Information Processing Systems 18* (Weiss, Y., Scholkopf, B., and Platt, J., eds.), pp. 507–514, Cambridge, MA : MIT Press, 2006.

[16] Cheung, Y. and Zeng, H. Local Kernel Regression Score for Selecting Features of High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering 21 , 12 (2009): 1798–1802.

[17] Mitra, P., Murthy, C. A., and Pal, S. K. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 24 , 3 (2002): 301–312.

[18] Hou, Y., Zhang, P., Yan, T., Li, W., and Song, D. Beyond Redundancies: A Metric-Invariant Method for Unsupervised Feature Selection. IEEE Transactions on Knowledge and Data Engineering 22 , 3 (2010): 348–364.

[19] Huang, Q., Jin, L., and Tao, D., An Unsupervised Feature Ranking Scheme by Discovering Biclusters. IEEE International Conference on Systems, Man and Cybernetics (SMC 2009), San Antonio, TX, USA, 11-14 October 2009.

[20] Li, Y., Sung, W.-K., and Miller, L. D., Multimodality as a Criterion for Feature Selection in Unsupervised Analysis of Gene Expression Data. Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), Los Alamitos, CA, 19-21 October 2005.

[21] Belkin, M. and Niyogi, P., Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. Advances in Neural Information Processing Systems 14, Vancouver, British Columbia, Canada, MIT Press, dec 2001.

[22] He, X. and Niyogi, P., Locality Preserving Projections. in *Advances in Neural Information Processing Systems 16* (Thrun, S., Saul, L., and Schölkopf, B., eds.), Cambridge, MA : MIT Press, 2004.

[23] Aksu, Y., Miller, D., Kesidis, G., and Yang, Q. Margin-Maximizing Feature Elimination Methods for Linear and Nonlinear Kernel-Based Discriminant Functions. IEEE Transactions on Neural Networks 21 , 5 (2010): 701–717.

[24] Ji, S. and Ye, J. Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection. IEEE Transactions on Neural Networks 19 , 10 (2008): 1768–1782.

[25] Peltonen, J. and Kaski, S. Discriminative Components of Data. IEEE Transactions on Neural Networks 16 , 1 (2005): 68–83.

[26] Casasent, D., Rozzi, W., and Fetterly, D. Projection synthetic discriminant function performance. Optical Engineering 23 (1984): 716–720.

[27] Ostrovsky, A. S., Mota, E. P., and Cuatianquiz, J. I. P. Optical Classification of Random Image fields using Spectral Synthetic Discriminant Functions. Optics and Lasers in Engineering 40 (2003): 43–53.

[28] Jing, X.-Y., Wong, H.-S., and Zhang, D. Face Recognition based on Discriminant Fractional Fourier Feature Extraction. Pattern Recognition Letters 27 (2006): 1465–1471.

[29] Wu, W., Walczak, B., Penninckx, W., and Massart, D. L. Feature reduction by Fourier Transform in Pattern Recognition of NIR data. Analytica Chimica Acta 331 (1996): 75-83.

[30] Goodman, J. W. Introduction to Fourier Optics. New York : Roberts & Company, third ed., 2004.

[31] Yu, C. and Neilson, D. Diffraction-Grating-based (de)Multiplexer using Image Plane Transformations. IEEE Journal of Selected Topics in Quantum Electronics 8 (nov/dec 2002): 1194 – 1201.

[32] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. Kernel Density Estimation via Diffusion. The Annals of Statistics 38 , 5 (2010): 2916–2957.

[33] Silverman, B., Some properties of a Test for Multimodality based on Kernel Density Estimates. in *Probability, Statistics and Analysis* (Kingman, J. F. C. and Reuter, G. E. H., eds.), pp. 248–259, Cambridge : Cambridge University Press, 1983.

[34] Hyvarinen, A. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10 , 3 (1999): 626–634.

[35] Duhamel, P. and Vetterli, M. Fast Fourier Transforms: A Tutorial Review and a State of the art. Signal Processing 19 , 4 (1990): 259–299.

[36] Newman D., Hettich S., Blake C., and Merz C. UCI Repository of Machine Learning Databases[online]. Available from: http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[37] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks. Nature Medicine 7 , 6 (2001): 673–679.

[38] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: Class Discovery and Class Prediction by gene Expression Monitoring. Nature Medicine 286 (1999): 531–537.

[39] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., Boer, M. L., Minden, M. D., Sallan, S., Lander, E., Golub, T., and Korsmeyer, S. J. MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a unique leukemia. Nature Medicine 30 (2001): 41–47.

[40] Zhang, Y. A MATLAB Package for Gene Selection[online]. Available from: http://users.cs.fiu.edu/∼yzhan004/genesel.html [2009].

[41] Calinskia, T. and Harabasza, J. A dendrite Method for Cluster Analysis. Communications in Statistics 3 , 1 (1974): 1–27.

[42] Padungweang, P. Supplement for A Discrimination Analysis for Unsupervised Feature Selection via Optic Diffraction Principle[online]. Available from: https://sites.google.com/site/praisan/supplementary-material [2011].

[43] Hyvarinen, A. and Oja, E. Independent Component Analysis: Algorithms and Applications. Neural Networks 13 , 4-5 (2000): 411–430.

[44] Zarzoso, V., Comon, P., and Phlypo, R. A Contrast Function for Independent Component Analysis Without Permutation Ambiguity. IEEE Transactions on Neural Networks 21 , 5 (2010): 863–868.

[45] Gavert, H., Hurri, J., Sarela, J., and Hyvarinen, A. FastICA[online]. Available from: http://www.cis.hut.fi/projects/ica/fastica/ [2007].

[46] Scott, D. W. On Optimal and Data-based Histograms. Biometrika 66 , 3 (1979): 605–610.

Appendix

APPENDIX

Grating diffraction pattern approximation

The total disturbance of diffraction pattern according to the light passing through the grating with G similar slits with the width of each slit is a with its center separated from each other by a distance of δ and $a \leq \delta$ can be expressed as follow:

$$
\begin{aligned}
E(q) = & \; c \int_{\xi} \sum_{g=0}^{G-1} \Pi_a(\xi - \xi_g) e^{-2\pi i q \xi} d\xi \\
= & \; c \int_{\xi} \Pi_a(\xi) e^{-2\pi i q \xi} d\xi + c \int_{\xi} \Pi_a(\xi - \delta) e^{-2\pi i q \xi} d\xi + c \int_{\xi} \Pi_a(\xi - 2\delta) e^{-2\pi i q \xi} d\xi \\
& + \ldots + c \int_{\xi} \Pi_a(\xi - (G-1)\delta) e^{-2\pi i q \xi} d\xi \\
E(q) = & \; c \int_{-\frac{a}{2}}^{\frac{a}{2}} e^{-2\pi i q \xi} d\xi + c \int_{\delta - \frac{a}{2}}^{\delta + \frac{a}{2}} e^{-2\pi i q \xi} d\xi + c \int_{2\delta - \frac{a}{2}}^{2\delta + \frac{a}{2}} e^{-2\pi i q \xi} d\xi + \ldots \\
= & \; c \left[- \frac{e^{-2\pi i q \xi}}{2\pi i q} \Big|_{-\frac{a}{2}}^{\frac{a}{2}} - \frac{e^{-2\pi i q \xi}}{2\pi i q} \Big|_{\delta - \frac{a}{2}}^{\delta + \frac{a}{2}} - \frac{e^{-2\pi i q \xi}}{2\pi i q} \Big|_{2\delta - \frac{a}{2}}^{2\delta + \frac{a}{2}} - \ldots \right] \\
= & \; \frac{c}{2\pi i q} \Big[-e^{-2\pi i q \frac{a}{2}} + e^{-2\pi i q (-\frac{a}{2})} - e^{-2\pi i q (\delta + \frac{a}{2})} + e^{-2\pi i q (\delta - \frac{a}{2})} \\
& - e^{-2\pi i q (2\delta + \frac{a}{2})} + e^{-2\pi i q (2\delta - \frac{a}{2})} - \ldots \Big] \\
= & \; \frac{c}{2\pi i q} (-e^{-2\pi i q \frac{a}{2}} - e^{-2\pi i q (\delta + \frac{a}{2})} - e^{-2\pi i q (2\delta + \frac{a}{2})} + e^{-2\pi i q (-\frac{a}{2})} \\
& + e^{-2\pi i q (\delta - \frac{a}{2})} + e^{-2\pi i q (2\delta - \frac{a}{2})} - \ldots) \\
= & \; - \frac{c}{2\pi i q} e^{-2\pi i q \frac{a}{2}} (1 + e^{-2\pi i q \delta} + e^{-2\pi i q 2\delta} + \ldots + e^{-2\pi i q (G-1)\delta}) \\
& + \frac{c}{2\pi i q} e^{2\pi i q \frac{a}{2}} (1 + e^{-2\pi i q \delta} + e^{-2\pi i q 2\delta} + \ldots + e^{-2\pi i q (G-1)\delta}) \\
= & \; \frac{c}{2\pi i q} (e^{2\pi i q \frac{a}{2}} - e^{-2\pi i q \frac{a}{2}}) (1 + (e^{-2\pi i q \delta}) + (e^{-2\pi i q \delta})^2 + (e^{-2\pi i q \delta})^3 \\
& + \ldots + (e^{-2\pi i q \delta})^{(G-1)})
\end{aligned}
$$

where

$$
\begin{aligned}
(e^{2\pi i q \frac{a}{2}} - e^{-2\pi i q \frac{a}{2}}) &= (\cos(\pi a q) + i \sin(\pi a q) - \cos(\pi a q) + i \sin(\pi a q)) \\
&= 2i \sin(\pi a q)
\end{aligned}
$$

and

$$(1+(e^{-2\pi iq\delta})+(e^{-2\pi iq\delta})^2+(e^{-2\pi iq\delta})^3+...+(e^{-2\pi iq\delta})^{(G-1)})=(\frac{e^{-2\pi iq\delta G}-1}{e^{-2\pi iq\delta}-1}).$$

Therefore,

$$\begin{aligned}
E(q) &= \frac{c}{2\pi iq}(2i\sin(\pi aq))(\frac{e^{-2\pi iq\delta G}-1}{e^{-2\pi iq\delta}-1}) \\
&= \frac{a}{a}\cdot\frac{c}{2\pi iq}(2i\sin(\pi aq))(\frac{-e^{-\pi iq\delta G}(e^{\pi iq\delta G}-e^{-\pi iq\delta G})}{-e^{-\pi iq\delta}(e^{\pi iq\delta}-e^{-\pi iq\delta})}) \\
&= ac(\frac{2i\sin(\pi aq)}{2i\pi aq})(e^{-\pi iq\delta(G-1)}(\frac{\sin(\pi q\delta G)}{\sin(\pi q\delta)})) \\
&= ae^{i(\phi r_0-\omega t)}(\frac{\sin(\pi aq)}{(\pi aq)})(e^{-i\alpha(G-1)}(\frac{\sin(\pi q\delta G)}{\sin(\pi q\delta)})) \\
&= a(e^{i(\phi r_0-\omega t-\alpha(G-1))}(\frac{\sin(\pi aq)}{(\pi aq)})(\frac{\sin(\pi q\delta G)}{\sin(\pi q\delta)}) \\
E(q) &= E_0(\frac{\sin(\pi aq)}{(\pi aq)})(\frac{\sin(G\pi q\delta)}{\sin(\pi q\delta)})
\end{aligned}$$

Consequently, the intensity function of the grating can be expressed as

$$P(q)=P_0\left(\frac{\sin(\pi aq)}{(\pi aq)}\right)^2\left(\frac{\sin(G\pi q\delta)}{\sin(\pi q\delta)}\right)^2$$

Biography

Praisan Padungweang was born in Chaiyaphum, Thailand, on April, 1979. He received B.Science. and M.Science., in Physics and Information Technology, from Khonkaen University, Thailand, in 2001 and 2006, respectively. His undergraduate study, from 1997 to 2000, was supported by the scholarship project of the promotion of science and mathematics talented teachers (PSMT.) from the institute for the Promotion of Teaching Science and Technology (IPST). In 2006, he has received a grant from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program under Grant No. Ph.D.1.O.CU/48/A.1. His field of interest includes various topics in Machine Learning including pattern analysis and image processing.

Education:

- Ph.D. Program in Computer Science, epartment of Mathematics, Faculty of Science, Chulalongkorn University, Bangkok, Thailand (October 2006 - September 2011).
- Visiting Ph.D. researcher in LINC laboratory at the Center for Advanced Computer Studies (CACS), University of Louisiana at Lafayette, United States (November 2010 - June 2011).
- M.Sc. in Information Techonology, Department of computer Science, Faculty of Science, Khonkean University, Khonkean, Thailand (May 2004 - March 2006).
- B.Sc. in Physics, Department of Physics, Faculty of Science, Khonkean University, Khonkean, Thailand (May 1997- March 2001).

Publication: Praisan Padungweang, Chidchanok Lursinsap and Khamron Sunat, “Univariate Filter Technique for Unsupervised Feature Selection Using a new Laplacian Score based Local Nearest Neighbors”, in *Proc. of 2009 Asia-Pacific Conference on Information Processing (APCIP 2009)*, Shenzhen, China, July 18-19, pp. 449-453, 2009.