

國立交通大學

電控工程研究所

博士論文

基於模糊線性區別分析之模糊分群法與結合空間
資訊之支撐向量機

A Clustering Algorithm Based on Fuzzy-Type Linear
Discriminant Analysis and
Spatial-Contextual Support Vector Machines

研究生：李政軒

指導教授：林進燈 教授

中華民國一〇一年一月

基於模糊線性區別分析之模糊分群法
與結合空間資訊之支撐向量機

A Clustering Algorithm Based on Fuzzy-Type Linear
Discriminant Analysis and
Spatial-Contextual Support Vector Machines

研 究 生：李政軒

Student: Cheng-Hsuan Li

指 導 教 授：林進燈

Advisor: Chin-Teng Lin



國立交通大學
電控工程研究所
博士論文

A Thesis

Submitted to the Institute of Electrical Control Engineering
College of Electrical and Computer Engineering

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor

in

Electrical and Computer Engineering

January 2012

Hsinchu, Taiwan, Republic of China

中華民國一〇一年一月

基於模糊線性區別分析之模糊分群法與結合空間資訊之支撐向量機

學生：李政軒

指導教授：林進燈 博士

國立交通大學電控工程研究所博士班

摘 要

統計學習演算法自動利用觀察資料來辨識複雜的樣本並進行決策。統計學習領域中有兩大主要議題：叢集分析與分類器設計。叢集分析演算法會將相似的樣本組織成同一個叢集；分類器則會利用現有的訓練樣本來決定新的未知樣本之類別。在本論文中，將提出模糊的分群演算法與融合空間資訊的分類器。在分群演算法方面，本文提出模糊線性區別分析之組間與組內分散矩陣，再搭配Fisher準則進行分群，此方法同時最小化群內資訊與最大化組間資訊。針對分類器的部分，透過空間資訊來調整支撐向量機的決策函數與限制式。利用真實資料的實驗結果顯示，本論文提出的方法可以有效地增加分群與分類的效能。

A Clustering Algorithm Based on Fuzzy-Type Linear Discriminant Analysis and Spatial-Contextual Support Vector Machines

Student: Cheng-Hsuan Li

Advisors: Dr. Chin-Teng Lin

Institute of Electrical Control Engineering
National Chiao Tung University

ABSTRACT

Statistical learning is trying to develop computer algorithms to recognize complex patterns and make decisions based on empirical data automatically. Two major issues are clustering and classification. Clustering organizes patterns into sensible clusters for patterns in the same cluster to be similar in a sense, whereas classification identifies the categories to which new patterns belong based on an available training set of data containing patterns of known categories. This thesis introduces a fuzzy-based clustering and a spatial-contextual classifier. Fuzzy-based clustering defines within- and between-cluster scatter matrices of a fuzzy-type linear discriminant analysis, and the clustering results are based on the Fisher criterion. The proposed clustering algorithm minimizes the within-cluster information and simultaneously maximizes the between-cluster information. For the classification part, a spatial-contextual term was used to modify the decision function and constraints of a support vector machine. Experimental results show that the proposed methods achieve good clustering and classification performance on famous real data sets.

誌 謝

首先誠摯的感謝指導教授林進燈博士、張志永博士與臺中教育大學教育測驗統計研究所郭伯臣博士，三位老師悉心的教導使我得以一窺特徵萃取、分群與分類等統計學習演算法領域的深奧，不時的討論並指點我正確的方向，使我在這些年中獲益匪淺。老師對學問的嚴謹更是我輩學習的典範。

四年半的日子裡，在工作上歷經逢甲大學應用數學系短期專任講師、臺中教育大學教育測驗統計研究所的行政助理，在生活上歷經結婚生子，變化頗大。但最令人懷念的是在實驗室裡共同的生活點滴，學術上的討論，撰寫論文的革命情感。感謝立偉學長於就學期間，不厭其煩的為我解惑，也感謝同學勝智在修課方面的幫忙，總是能帶來課堂與所務的第一手消息。實驗室的鈞翔、志勝、士勛等學弟妹們當然也不能忘記，你們的幫忙我銘感在心。

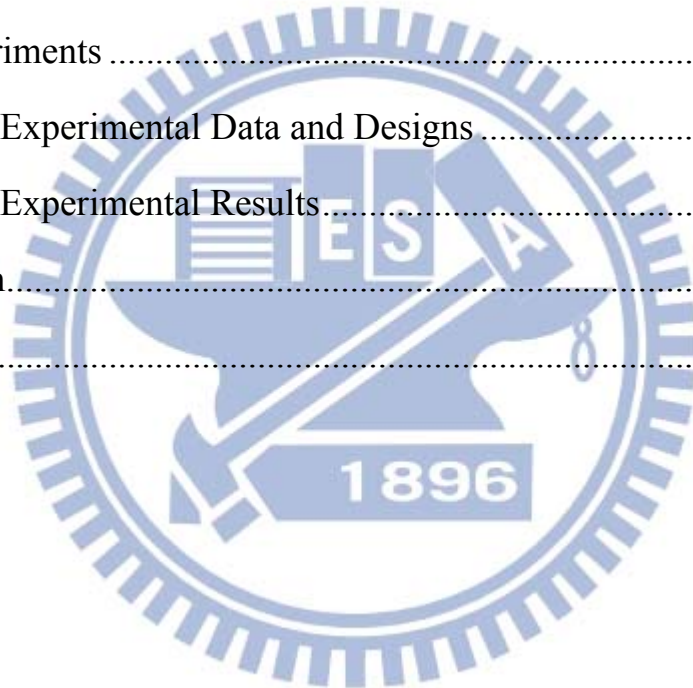
老婆省華在背後的默默支持更是我前進的動力，沒有省華的體諒、包容，相信這兩年的生活將是很不一樣的光景。女兒沛芩的笑容，是最後一個學期生活上最棒的潤滑劑，總能在我忙碌的工作之餘，放鬆我的心情，讓我能夠重啟熱情迎接每一忙碌的日子。

最後，謹以此文獻給我摯愛的父親、母親與媽咪。

Contents

Chinese abstract	i
English abstract	ii
Acknowledgement.....	iii
Contents.....	iv
List of tables	vi
List of figures	viii
List of symbols.....	xii
1. Introduction.....	1
2. Literature Review of Fuzzy-based Clustering Algorithms.....	10
2.1 Fuzzy C-means Clustering Algorithm.....	10
2.2 Gustafson-Kessel algorithm.....	11
2.3 Fuzzy Compactness and Separation	13
2.4 Other FCM-type Clustering Algorithms.....	15
3. LDA-based Clustering Algorithm.....	16
3.1 Review of LDA.....	16
3.2 FLDC Algorithm.....	17
3.3 Experiments	20
3.3.1 Experimental Data and Designs	20
3.3.2 Experimental Results.....	23
4. The Support Vector Machine and Its Spectral-Spatial Classification Schemes.....	34
4.1 Support Vector Machine.....	34

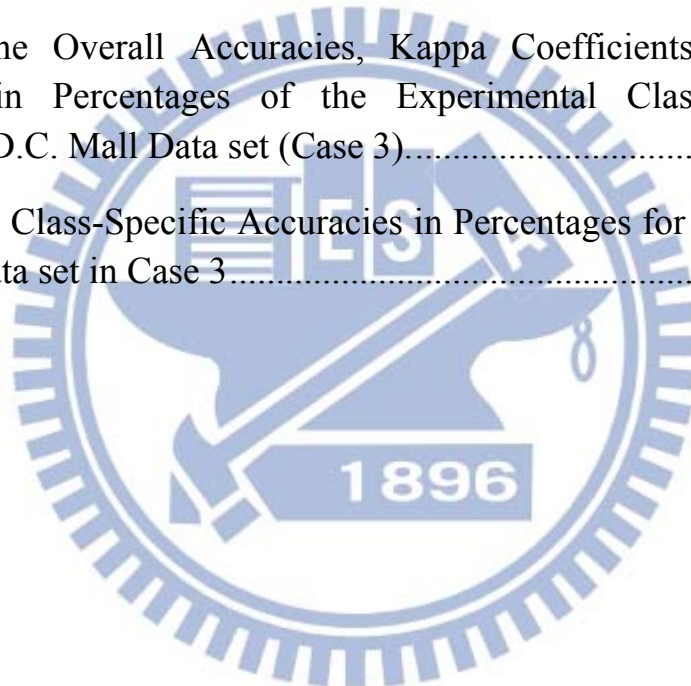
4.2 Spectral-Spatial Classification Scheme Based on Partitional Clustering Techniques	37
4.3 Context-Sensitive Semi-supervised SVM	40
5. Spatial-Contextual Support Vector Machines	43
5.1 A Spatial-contextual Support Vector Machine in the Original Space	43
5.2 A Spatial-Contextual Support Vector Machine in the Feature Space	49
5.3 Classification System of SCSVM and SCSVMF	50
5.4 Experiments	52
5.4.1 Experimental Data and Designs	52
5.4.2 Experimental Results.....	56
6. Conclusion.....	75
References	78



List of tables

Table 1 Descriptions of Three Real Data Sets	23
Table 2 The Mean, Standard Deviation, Maximum, and Minimum Accuracy of Clustering for Three Real Data sets.	33
Table 3 The Mean, Standard Deviation, Maximum, and Minimum Accuracy of Clustering for Three Real Data sets of FMSFA, Where LD Represents the Latent Dimension	33
Table 4 Sixteen Categories and Corresponding Number of Pixels in the Indian Pine Site Image	53
Table 5 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in the IPS Data set.....	58
Table 6 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the IPS Data set	58
Table 7 The Class-specific Accuracies in Percentages for the IPS Data set	59
Table 8 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in Washington D.C. Mall Data set (Case 1).....	64
Table 9 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in Washington D.C. Mall Data set (Case 2).....	66
Table 10 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in Washington D.C. Mall Data set (Case 3).....	67

Table 11 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the Washington D.C. Mall Data set (Case1).....	68
Table 12 The Class-Specific Accuracies in Percentages for the Washington D.C. Mall Data set in Case 1.....	68
Table 13 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the Washington D.C. Mall Data set (Case 2).....	68
Table 14 The Class-Specific Accuracies in Percentages for the Washington D.C. Mall Data set in Case 2.....	69
Table 15 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the Washington D.C. Mall Data set (Case 3).....	69
Table 16 The Class-Specific Accuracies in Percentages for the Washington D.C. Mall Data set in Case 3.....	70



List of figures

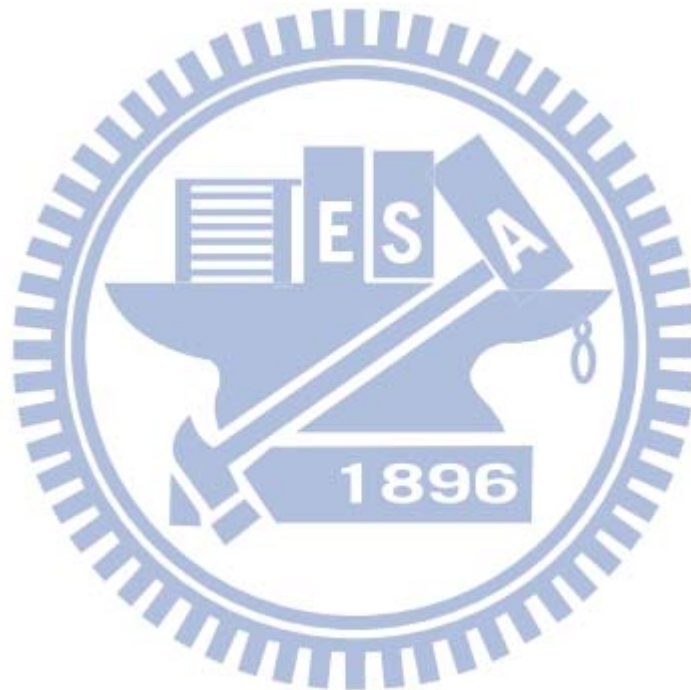
Figure 1. The spectral values obtained from the Indian Pine Site data set. The purple represents the Soybeans-min till patterns and the yellow represents the Corn-no till patterns. These two classes have similar spectral properties.....	5
Figure 2. The support vector machine (SVM) classification results of the Indian Pine Site image, containing speckle-like errors.....	6
Figure 3. (a) The “x” remarks 50 random samples chosen from the multivariate normal distribution with mean $[0.5,0]^T$ and covariance $\begin{bmatrix} 0.8 & 0.7 \\ 0.7 & 0.8 \end{bmatrix}$, and the “o” remarks 50 random samples chosen from the multivariate normal distribution with mean $[-0.5,0]^T$ and covariance $\begin{bmatrix} 0.8 & -0.7 \\ -0.7 & 0.8 \end{bmatrix}$; (b) The clustering results of (a) applying FCM; (c) the clustering results of (a) applying GK algorithm.	12
Figure 4. Ten artificial data sets [53]-[54] were used in this study. The first three data sets were generated with 10 additional noise features. The number of clusters appears in parentheses.....	22
Figure 5. The results of clustering the “Four gauss” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	26
Figure 6. The results of clustering the “Easy doughnut” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.....	27
Figure 7. The results of clustering the “Difficult doughnut” data set using twelve clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	28
Figure 8. The results of clustering the “Boat” data set using twelve clustering algorithms. The best clustering results from applying GG and	

GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	29
Figure 9. The results of clustering the “Noisy lines” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	30
Figure 10. The results of clustering the “Petals” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	30
Figure 11. The results of clustering the “Saturn” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	31
Figure 12. The results of clustering the “Regular” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.	32
Figure 13. Flowchart of the SVM+EM [31].	37
Figure 14. Example of SVM+EM classification [31].	39
Figure 15. The left and right images represent the first-order and second-order neighborhood systems in the original space, respectively.	41
Figure 16. Example of training and related context patterns in the feature space [32].	41
Figure 17. The pixels enclosed by bold lines represent the first-order neighborhood system used in SCSVM.	43
Figure 18. An example of the spatial-contextual information with the second-order neighborhood system of pattern x_i in the original space.	45
Figure 19. The left panel shows the decision boundary (solid black line) obtained by SVM. The center panel shows the semi-labels of the patterns in	

the second-order neighborhood system of x_j . The right panel shows the decision boundary (solid red line) obtained of SCSVM.....	46
Figure 20. A multiclass case of the spatial contextual information defined by the OAO strategy (class 1 versus class 2) for pattern x_i in the neighborhood system ∂x_i^o . The labels of class 1 and class 2 are defined as +1 and -1, respectively.....	47
Figure 21. A multiclass case of the spatial contextual information defined by the OAA strategy (class 1 versus all others) for pattern x_i in the neighborhood system ∂x_i^o . The label of class 1 is defined as +1 and the labels of the remaining classes (class 2 and class 3) are defined as -1.....	48
Figure 22. SCSVM and SCSVMF classification systems.....	51
Figure 23. A portion of the Indian pine site image measuring 145×145 pixels.....	52
Figure 24. The ground truth of the Indian pine site data set.....	52
Figure 25. The false-color IR image of a portion of Washington D.C. Mall image measuring 205×307 pixels. There are seven categories: grass, tree, roof, water, road, trail, and shadow.....	54
Figure 26. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the IPS data set.....	57
Figure 27. The classification maps of the IPS data set by the highest performance of each type classifier.....	62
Figure 28. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the Washington D.C. Mall data set in case 1.....	64
Figure 29. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the Washington D.C. Mall data set in case 2.....	66
Figure 30. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the Washington D.C. Mall data set in case 3.....	67

Figure 31. The classification maps of a portion of the Washington D.C. data set (case 3) by the highest performance of each type classifier. 73

Figure 32. The classification maps of a portion of the Washington D.C. data set (case 3) of SCSVM (OAO) and SCVM (OAA) with $M=4$ and different parameters $\gamma=0, 0.1, \text{ and } 0.3$ 74



List of symbols

- L : the number of clusters or classes
- N : the number of samples
- n : the number of training samples
- M : the number of samples in the neighborhood system
- X : A hyperspectral d -dimensional image of size $I \times J$ pixels
- D : the training data set
- H_i : the set of samples in class i
- H : a Hilbert space
- N_i : the number of samples in class i
- R : the set of real numbers
- R^d : the d dimensional Euclidean space
- R^+ : the set of positive real numbers
- ∂x_i^O : a neighborhood system w.r.t. x_i in the original space
- ∂x_i^F : a neighborhood system w.r.t. x_i in the feature space
- x : an unlabeled pattern
- x_j : the j -th sample
- $x_j^{(i)}$: the j -th sample in class i
- \bar{x}_{ij} : the j -th sample in the neighborhood system ∂x_i
- c_i : the center of the cluster i or class i
- c : the total mean of samples
- v_s : the s -th large eigenvector

- w : a normal vector to the decision hyperplane of support vector machine
- ξ : the vector whose elements are slack variables of the support vector machine
- ψ_{ij} : the vector whose elements are slack variables of the context-sensitive semi-supervised support vector machine
- κ_{ij} : the weights of the importances of context patterns in the context-sensitive semi-supervised support vector machine
- α : the vector with elements that are Lagrange multipliers
- d : the dimensionality of the original space
- m : the weighting exponent
- p : the dimensionality of the reduced space
- r : the regularization parameter of LDA-based clustering
- λ_s : the s -th large eigenvalue
- u_{ij} : the membership grade of the j -th sample in cluster i
- η_i : the tradeoff parameter of fuzzy compactness and separation
- β : the parameter to control η_i
- y_i : the class label w.r.t. the training sample x_i
- y_{ij} : the semi-label of \bar{x}_{ij}
- b : a constant to control the decision hyperplane of the support vector machine
- ξ_i : a slack variable of the support vector machine w.r.t. x_i
- C : a penalty parameter of the support vector machine
- α_i : a Lagrange multiplier
- γ : a nonnegative parameter that controls the effect of spatial-contextual information

- S_{FW} : the fuzzy within-cluster scatter matrix
 S_{FB} : the fuzzy between-cluster scatter matrix
 S_w^{LDA} : the within-class scatter matrix of linear discriminant analysis
 S_b^{LDA} : the between-class scatter matrix of linear discriminant analysis
 S_w^{UFLDA} : the within-class scatter matrix of LDA-based clustering
 S_b^{ULDA} : the between-class scatter matrix of LDA-based clustering
 K : a kernel matrix
 Σ_i : a positive semi-definite matrix for computing an adaptive distance norm of Gustafson-Kessel algorithm
 F_i : the fuzzy covariance matrix of Gustafson-Kessel algorithm
 J_{FCM} : the cost function of fuzzy c-means
 J_{GK} : the cost function of Gustafson-Kessel algorithm
 J_{FCS} : the cost function of fuzzy compactness and separation
 J_{LDA} : the objective function of linear discriminant analysis
 J_{FLDC} : the objective function of LDA-based clustering
 ϕ : a nonlinear feature mapping
 κ : a kernel function
 f_{SVM} : the decision function of the support vector machine
 f_{SCSVM} : the decision function of the spatial-contextual support vector machine
 m^+ : the function with the output being the number of pixels in the neighbor system belonging to class +1
 m^- : the function with the output being the number of pixels in the neighbor system belonging to class -1

1. Introduction

Researchers have developed numerous statistical learning algorithms for applications in various areas of science, finance, and industry in recent years. Statistical learning comprises several different paradigms such as classification, regression, feature extraction, dimensionality reduction and density estimation [3]. The basic idea of classification methods for feature space data is to partition up the entire feature space into L exhaustive, nonoverlapping regions, where L is the number of classes present in the scene, so that every point in the feature space is uniquely associated with one of the L classes [22].

The classification algorithms can be divided into two main categories according to the learning process. Supervised classification, or simply classification, is the learning process of inferring a function to classify unknown patterns using the training data to train the rule [66], i.e., a set of training samples is available and the classifier exploits this a priori known information [2].

The other type of learning process is called unsupervised classification, or simply clustering. It is referred to as unsupervised because it does not use training samples [22]. Clustering assesses the relationships among samples of a data set by organizing the patterns into different groups. After clustering, patterns in one group show greater similarity to each other than those belonging to different groups without any prior known information [1]. Clustering analysis can detect underlying structures within data, for classification and pattern recognition, and for model reduction and optimization [2], [4]-[5].

Clustering algorithms are most commonly used as an aid to selecting a class list and training samples for the classes in that list. That is, clustering

may be a means of preprocessing the data for a supervised classification procedure. A clustering scheme may be applied to the data for each class separately and representative samples for each group within the class used as the prototypes for that class [66]. Fundamentally, to be optimally useful, a classification must have classes that are (simultaneously) “of information value, exhaustive, and separable.” The training samples for supervised learning generally are selected with emphasis on the former one. Clustering is a useful tool of the training process to achieve the latter two. It can be a useful procedure, though, in defining spectral classes and training for them by breaking up the distribution of pixels in feature space into subunits so that one can observe what is likely to be separable from what. It allows one to locate the prevailing modes in the feature space, if any prevalence exists [22].

Recent statistical learning algorithms [17]-[19] use both labeled and unlabeled samples for training. These algorithms are called semi-supervised learning process, and fall between unsupervised pattern recognition and supervised recognition. The aim of this thesis is to develop an unsupervised clustering algorithm and a semi-supervised classification algorithm. The former one is a fuzzy-based clustering which considers both within- and between-information of clusters, and the latter one is a semi-supervised classification algorithm which takes into account both spectral and spatial information.

Fuzzy-based clustering, which determines if a vector belongs to a specific cluster to a certain degree, have been the subject of intensive research in the past three decades [2], [4]-[8]. Fuzzy c-means (FCM) clustering is one of the most well-known clustering methods [7]-[8], and researchers have developed many advanced FCM-type clustering algorithms. The Gustafson-Kessel (GK) algorithm [9] is a well-known

algorithm in this category. This algorithm employs an adaptive distance norm to detect clusters of different geometrical shapes in one data set [2]. Krishnapuram and Keller [52] proposed a new clustering model, called possibilistic c-means (PCM), which relaxes the following constraint: “the sum of the membership values of every sample to all clusters is 1.” This approach avoids the outliers belonging to one or more clusters. In 1997, the fuzzy-possibilistic c-means (FPCM) [10] was proposed to generate both possibility and membership values. However, the possibility values generated by FPCM become very small as the size of the data set increases. To eliminate the problem of FPCM and take advantage of the benefits of FCM and PCM, the possibilistic fuzzy c-means (PFCM) was proposed in 2005 [11].

Some FCM-type algorithms, such as the Gath-Geva (GG) algorithm, employ an adaptive distance norm based on the fuzzy maximum likelihood estimates [5], [12]. Chatzis and Varvarigou [13] proposed a robust fuzzy clustering algorithm based on the fuzzy treatment of finite mixtures of multivariate Student’s t -distributions (FSMM). This approach uses finite mixtures of multivariate Student’s t distributions instead of finite Gaussian mixture models (GMMs). Chatzis and Varvarigou [56] combined the advantages of factor analysis and proposed a fuzzy mixture of Student’s t factor analyzers (FMSFA). FMSFA provides a well-established observation space dimensionality reduction framework for fuzzy clustering algorithms based on factor analysis. This simultaneously achieves fuzzy clustering and a reduction in local dimensionality within each cluster. Their experimental results show that FMSFA outperforms finite mixtures of Student’s t -factor analyzers (t MFA) [57], a modification of the fuzzy c-varieties algorithm with regularization by Kullback–Leibler information (KLFCV) [58], and the mixture of factor analyzers (MFA) model [59].

Most fuzzy-based clustering algorithm by minimizing a cost function, only based on the sum of distances between samples to their cluster centers [2], which is equal to the trace of the within-cluster scatter matrix [14]-[15]. Researchers have recently used linear discriminant analysis (LDA) [14] for dimensional reduction in supervised classification problems. LDA uses the mean vector and covariance matrix of each class to formulate within-class, between-class, and mixture-class scatter matrices. Two similar fuzzy-based clustering algorithms based on fuzzy within-cluster, between-cluster, and total scatter matrices are proposed in [15] and [16]. The objective function of fuzzy compactness and separation (FCS) [15] is based on the difference of fuzzy within- and between-cluster scatter matrices. This minimizes the measurement of compactness, but simultaneously maximizes the separation measure. However, the within- and between-class scatter matrices of LDA are not the special case of the proposed fuzzy within- and between-cluster scatter matrices in the supervised learning problem. Moreover, based on the Fisher criterion, the LDA method finds features such that the ratio of the between-class scatter to the average within-class scatter is maximized in a lower dimensional space. Of the concept of class scattering to class separation, the Fisher criterion takes the large values from samples when they are well clustered around their mean within each class, and the clusters of the different classes are well separated [2]. The Fisher criterion is formulated as a function of class statistics. For these reasons, this thesis proposes a clustering algorithm based the Fisher criterion [4].

The first part of the thesis is to propose a fuzzy-based clustering which is based on the fuzzy-based within- and between-cluster scatter matrices. In addition, the Fisher criterion is used to form the objective function. This means that the proposed clustering algorithm take into account not only the within- and between-information of the distribution of data but also the

interaction of the within- and between-information. Chapters 2-3 present the fuzzy-based clustering algorithm. Chapter 2 introduces some recently proposed fuzzy-based clustering algorithms. Chapter 3 details the proposed clustering algorithm based on both within- and between-cluster scatter matrices, extended from linear discriminant analysis (LDA) [4].

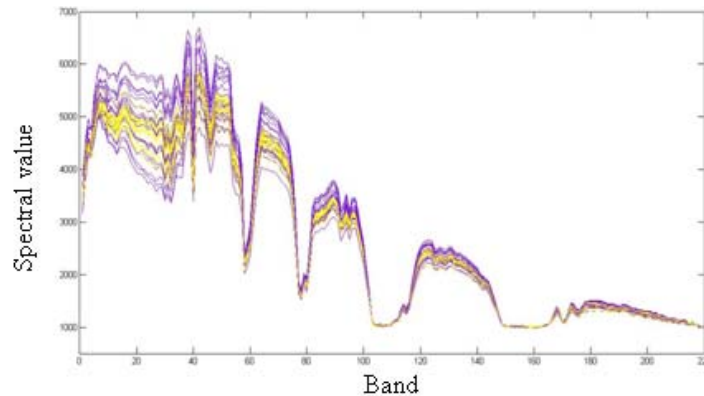


Figure 1. The spectral values obtained from the Indian Pine Site data set. The purple represents the Soybeans-min till patterns and the yellow represents the Corn-no till patterns. These two classes have similar spectral properties.

In hyperspectral image classification, spectral-domain based classifiers often lead to imprecise estimation of different land-cover classes that have very similar spectral properties, which makes it difficult to distinguish unlabeled patterns [20]-[21]. Fig. 1 shows the spectral values obtained from patterns of two categories in the Indian Pine Site data set: Soybeans-min till (purple color) and Corn-no till (yellow color) [22]. These two different classes have very similar spectral properties. Hence, employing these classes to train conventional classifiers (e.g., maximum likelihood classifier (ML) [2], [15], k -nearest neighbor classifier (k -NN) [2], [15], and support vector machine (SVM) [23]-[24]) leads to poor classification performance, producing a speckle-like classification map [20]-[21], [25]. Fig. 2 shows

that the support vector machine (SVM) classification map of Indian Pine Site includes a number of speckle-like errors.



Figure 2. The support vector machine (SVM) classification results of the Indian Pine Site image, containing speckle-like errors.

Considering both spectral and spatial-contextual information, using a semi-supervised learning algorithm is an effective way to decrease speckle-like errors when interpreting a hyperspectral image. There are two main methods for combining spectral and spatial-contextual information. The graph-based technique [18]-[19], [26]-[32] uses the typical method of performing a regularization in which “similar” features belong to the same class. This method associates the vertices of a graph with the complete set of samples, and then builds the regularization depending on the variables defined on the vertices [18]. The other approach is to use fixed-window-based methods, such as Markov random fields [20]-[21], morphological filtering [28], or morphological leveling [29]-[30]. This approach improves the classification performance of hyperspectral images compared to pixel-wise methods [31].

Jackson and Landgrebe [20] applied a Gaussian function to the Bayesian decision rule with Markov random fields (MRF), Bayesian contextual classifier based on MRF (ML_MRF), to mitigate the

speckle-like errors. Their method achieves improved performance in classification maps. Another study suggests applying similar concepts to develop a MRF-based k -nearest neighbors classifier and Parzen classifier [21]. However, MRF-based classifiers are still constrained by statistical estimation (e.g., the covariance matrix of ML based on a Gaussian distribution) or the amount of learning data.

The support vector machine [23] is a pattern classification technique proposed by Vapnik et al. Unlike traditional methods, which minimize empirical training errors, SVM attempts to minimize the upper bound of the generalization error by maximizing the margin between the separating hyperplane and the training data. Hence, SVM is a distribution-free algorithm that can overcome the problem of poor statistical estimation. SVM also achieves greater empirical accuracy and better generalization capabilities than other standard supervised classifiers [3] [34]-[35]. In particular, SVM performs well for high-dimensional data classification with a few training samples [37]-[38], and is robust to the Hughes phenomenon [32]-[33], [35], [37]-[38].

Moreover, many studies [30]-[33] show that support vector machines with both spectral and spatial information achieve effective and stable hyperspectral image classification. A context-sensitive semi-supervised support vector machine (CS⁴VM) [32] uses the context of neighborhood patterns as semi-patterns to solve the problem of noisy training patterns. In this case, noisy training patterns are mislabeled patterns that introduce distorted information to a classifier. CS⁴VM is a semi-learning approach in which the computational cost increases as the number of semi-samples increases.

Tarabalka et al. [31] presented a spectral-spatial classification scheme based on partitional clustering techniques (SVM+EM). This approach

segments an image into more homogeneous regions and combines the results of these regions using pixel-wise SVM classification. A spatial post-regularization (PR) of the classification map reduces the noise. This approach is particularly suitable for classifying images with large spatial structures, when spectral responses of different classes are dissimilar, and when classes contain a comparable number of pixels. If the spectral responses are not significantly different, this approach may result in misclassification [31].

The second part of this thesis uses two neighborhood systems, that one is in the original space and the other one is in the feature space, to modify the constrain and decision rule of the support vector machine, and proposes a spatial-contextual support vector machine to overcome the speckle-like errors. Chapters 4-5 focus on the spectral-spatial classification schemes. Chapter 4 introduces the SVM and some recently spectral-spatial classification algorithms. Chapter 5 describes two spatial-contextual support vector machine classification algorithms (SCSVMs) [39] that modifies the decision function and constraints of a support vector machine (SVM) using a spatial-contextual term in the original space or in the feature space, which are based on the concept of the Markov random fields in the original space or k -nearest neighborhoods in the feature space, respectively.

The thesis is devoted to fuzzy-based clustering algorithm, fuzzy linear discriminant clustering (FLDC), and semi-supervised image classification, spatial-contextual support vector machine. First, in Chapter 3, fuzzy-based within- and between-cluster scatter matrices extended from the within- and between-class scatter matrices of LDA are introduced. Furthermore, the Fisher criterion composed by the fuzzy-based scatter matrices is used to form the objective function. FLDC considers not only the within- and between-information of the data distribution but also the interaction of the

within- and between-information. The results of experiments on both synthetic and real data show that the proposed clustering algorithm can generate similar or better clustering results than eleven popular clustering algorithms: K-means, K-medoid, FCM, the Gustafson-Kessel, Gath-Geva, possibilistic c-means, fuzzy-possibilistic c-means, possibilistic fuzzy c-means, fuzzy compactness and separation, a fuzzy clustering algorithm based on a fuzzy treatment of finite mixtures of multivariate Student's- t distributions algorithms, and a fuzzy mixture of Student's t factor analyzers model.

Then, in Chapter 5, two neighborhood systems is used to overcome the similar spectrum problem in support vector machine. Two semi-supervised classifiers, spatial-contextual support vector machines (SCSVMs), are proposed by modifying the constrain and the decision function of support vector machine. To evaluate the effectiveness of SCSVM, the experiments in this study compare the performances of other classifiers: a support vector machine (SVM), context-sensitive semi-supervised support vector machine (CS4VM), maximum likelihood classifier (ML), Bayesian contextual classifier based on Markov random fields (ML_MRF), and k -nearest-neighbor classifier (k -NN). Experimental results show that the proposed method achieves good classification performance on famous hyperspectral images (the Indian Pine site and the Washington, D.C. Mall data sets). The overall classification accuracy of for the hyperspectral image of the Indian Pine site dataset with 16 classes is 95.5%. The kappa accuracy is up to 94.9%, and the average accuracy of each class is up to 94.2%.

2. Literature Review of Fuzzy-based Clustering Algorithms

The aim of clustering algorithms is to identify unknown data structures, such as natural groups or clusters, by measuring the similarities between samples. The samples within a cluster or group are more similar to each other than those pixels belonging to other clusters [3], [40]. This section reviews some well-known fuzzy-based clustering algorithms.

2.1 Fuzzy C-means Clustering Algorithm

Fuzzy c-mean clustering (FCM) is the fuzzy equivalent of the nearest mean “hard” clustering algorithm [1]-[2], [5]-[6], [41], and minimizes the cost function

$$J_{FCM}(u_{ij}, c_i) = \sum_{i=1}^L \sum_{j=1}^N (u_{ij})^m \|x_j - c_i\|^2$$

with respect to membership grade u_{ij} and c_i , the center of fuzzy cluster i , where $x_j \in R^d$, N is the number of samples, $L > 1$ is the number of clusters, and $m \in (1, \infty)$ is a weighting exponent.

The FCM algorithm assigns the memberships to x_j . These memberships are inversely related to the relative distance of x_j to the L cluster centers $\{c_i\}$. The formulation of criterion J_{FCM} could be regarded as the trace of the fuzzy within-cluster scatter matrix S_{FW} [2], which is defined as

$$S_{FW} = \sum_{i=1}^L \sum_{j=1}^N (u_{ij})^m (x_j - c_i)(x_j - c_i)^T.$$

Equation above is similar to the within-class scatter matrix of LDA in that this criterion only considers the within-cluster scatter matrix. A consideration the within-cluster similarity is the only criterion. Based on previous suggestions [34], the division into clusters should be characterized by within-cluster similarity and between-cluster (external) dissimilarity. This is the reason why this study applies the Fisher criterion.

2.2 Gustafson-Kessel algorithm

The Gustafson-Kessel (GK) algorithm [9] is a well-known example of FCM-type clustering algorithms. The GK algorithm employs an adaptive distance norm to detect clusters of different geometrical shapes in one data set [5]. FCM is suitable for clusters with similar distributions. If clusters with very different distributions like Fig. 3(a), the “x” remarks 50 random samples chosen from the multivariate normal distribution with mean $[0.5,0]^T$ and covariance $\begin{bmatrix} 0.8 & 0.7 \\ 0.7 & 0.8 \end{bmatrix}$, and the “o” remarks 50 random samples chosen from the multivariate normal distribution with mean $[-0.5,0]^T$ and covariance $\begin{bmatrix} 0.8 & -0.7 \\ -0.7 & 0.8 \end{bmatrix}$, the clustering results of FCM (Fig. 3(b)) are frequently wrong, especially, on the left-bottom part of cluster 1 and the right-bottom part of cluster 2. The Gustafson-Kessel (GK) algorithm defines the fuzzy covariance matrices, which are used to compute generalized squared Mahalanobis distances, to solve this problem. Fig. 3(c) shows the clustering results of GK algorithm which is more similar to Fig. 3(a) than FCM. That is, the GK algorithm can detect clusters of different geometrical shapes in one data set.

The objective function of GK algorithm [9] is defined as

$$J_{GK}(u_{ij}, c_i, \Sigma_i) = \sum_{i=1}^L \sum_{j=1}^N (u_{ij})^m (\mathbf{x}_j - \mathbf{c}_i)^T \Sigma_i (\mathbf{x}_j - \mathbf{c}_i)$$

where the matrices Σ_i , which adapt the distance norm to the local topological structure of the data [5], serve as optimization variables. Since Σ_i should be a positive definite matrix, the common approach is to constrain the determinant of Σ_i (i.e., $\det(\Sigma_i) = \rho_i$, $\rho_i > 0$, $i = 1, \dots, L$).

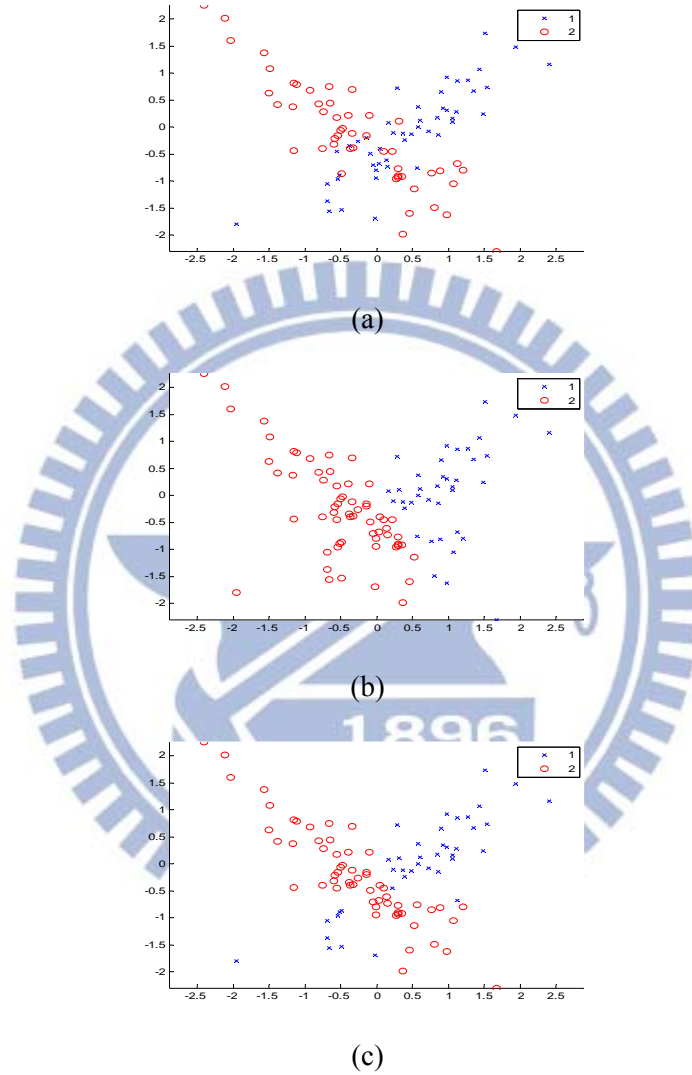


Figure 3. (a) The “x” remarks 50 random samples chosen from the multivariate normal distribution with mean $[0.5, 0]^T$ and covariance $\begin{bmatrix} 0.8 & 0.7 \\ 0.7 & 0.8 \end{bmatrix}$, and the “o” remarks 50 random samples chosen from the multivariate normal distribution with mean $[-0.5, 0]^T$ and covariance $\begin{bmatrix} 0.8 & -0.7 \\ -0.7 & 0.8 \end{bmatrix}$; (b) The clustering results of (a) applying FCM; (c) the clustering results of (a) applying GK algorithm.

Using the Lagrange multiplier method, Σ_i is obtained by

$$\Sigma_i = (\rho_i \det(F_i))^{1/d} F_i^{-1},$$

where F_i is the fuzzy covariance matrix [5], [9] of the i -th cluster defined by:

$$F_i = \frac{\sum_{j=1}^N (u_{ij})^m (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T}{\sum_{j=1}^N (u_{ij})^m}.$$

2.3 Fuzzy Compactness and Separation

Previous studies [15], [16] have proposed two similar fuzzy-based clustering algorithms based on fuzzy within-cluster, between-cluster, and total scatter matrices. The objective function of the fuzzy compactness and separation (FCS) [15] is based on fuzzy between- and within-cluster scatter matrices. This approach minimizes the measurement of compactness, and simultaneously maximizes the separation measure.

The fuzzy between-cluster scatter matrix S_{FB} and within-cluster scatter matrix S_{FW} are defined as

$$S_{FB} = \sum_{i=1}^c \sum_{j=1}^n \eta_i(u_{ij})^m (\mathbf{x}_j - \mathbf{c})(\mathbf{x}_j - \mathbf{c})^T$$

and

$$S_{FW} = \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^m (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T$$

where $\mathbf{c} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$. The objective function of FCS is defined as

$$\begin{aligned} J_{FCS}(\mathbf{u}_{ij}, \mathbf{c}_i) &= \text{tr}(S_{FW}) - \text{tr}(S_{FB}) \\ &= \sum_{i=1}^L \sum_{j=1}^N (u_{ij})^m \|\mathbf{x}_j - \mathbf{c}_i\|^2 - \sum_{i=1}^L \sum_{j=1}^N \eta_i(u_{ij})^m \|\mathbf{x}_j - \mathbf{c}\|^2. \end{aligned}$$

By minimizing J_{FCS} , the proposed method uses the following equations to mutually update each other:

$$u_{ij} = \frac{(\|x_j - c_i\|^2 - \eta_i \|c_i - c\|^2)^{-1/(m-1)}}{\sum_{k=1}^L (\|x_j - c_k\|^2 - \eta_k \|c_k - c\|^2)^{-1/(m-1)}}$$

and

$$c_i = \frac{\sum_{j=1}^N (u_{ij})^m x_j - \eta_i \sum_{j=1}^n (u_{ij})^m c}{\sum_{j=1}^N (u_{ij})^m - \eta_i \sum_{j=1}^n (u_{ij})^m},$$

where the parameter η_i could be set up with

$$\eta_i = \frac{(\beta/4) \min_{i' \neq i} \|c_i - c_{i'}\|^2}{\max_k \|c_k - c\|^2},$$

and $\beta \in [0,1]$ is the parameter to be pre-determined. The objective function proposed by Yin et al. [43] is a special case of FCS in which the parameters η_i are all set to $1/(L(L-1))$.

The Fisher criterion, the trace of the product of the inverse of the within-class scatter matrix and the between-class scatter matrix, takes large values when samples are well clustered, around their mean within each class, and the clusters of the different classes are well separated [2]. This approach is widely used in different applications [42], [43]-[44]. The following discussion introduces new definitions of unsupervised cluster scatter matrices. The corresponding objective function is based on the Fisher criterion including the interaction of cluster scatter matrices.

2.4 Other FCM-type Clustering Algorithms

Krishnapuram and Keller [52] proposed a new clustering model, called possibilistic c-means (PCM), that relaxes a constraint (“the sum of the membership values of every sample to all clusters is 1”) to interpret the membership function or degree of typicality in a possibilistic sense [45]. The fuzzy-possibilistic c-means (FPCM) [10] was proposed in 1997 to generate both possibility and membership values. However, the possibility values generated by FPCM become very small as the size of the data set increases. To eliminate the problem of FPCM and take advantage of the benefits of FCM and PCM, the possibilistic fuzzy c-means (PFCM) was proposed in 2005 [46].

Some FCM-type algorithms, such as the Gath-Geva (GG) algorithm, employ an adaptive distance norm based on the fuzzy maximum likelihood estimates [2], [30]. Chatzis and Varvarigou [27] proposed a robust fuzzy clustering algorithm based on a fuzzy treatment of finite mixtures of multivariate Student’s- t distributions (FSMM). This approach uses finite mixtures of multivariate Student’s t distributions instead of finite Gaussian mixture models (GMMs).

3. LDA-based Clustering Algorithm

This chapter introduces a novel clustering algorithm, called fuzzy linear discriminant clustering (FLDC), that accounts both within- and between-cluster information [4]. Since the scatter matrices are extended from the LDA, Section 3.1 reviews the LDA.

3.1 Review of LDA

LDA is often used for dimension reduction in classification problems. Because it uses the mean vector and covariance matrix of each class, LDA is often referred to as the parametric feature extraction method [14]. Within-class, between class, and mixture scatter matrices are frequently used to formulate the criterion of class separability.

Suppose that $H_i = \{x_1^{(i)}, \dots, x_{N_i}^{(i)}\} \subset R^d$ are the set of samples in class i , N_i is the number of samples in class i , $i = 1, \dots, L$, and $N = N_1 + \dots + N_L$ is the number of all training samples. LDA defines the between-class scatter matrix S_b^{LDA} and the within-class scatter matrix S_w^{LDA} as

$$S_b^{LDA} = \sum_{i=1}^L \frac{N_i}{N} (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T$$

and

$$S_w^{LDA} = \sum_{i=1}^L \sum_{j=1}^{N_i} \frac{1}{N} (\mathbf{x}_j^{(i)} - \mathbf{c}_i)(\mathbf{x}_j^{(i)} - \mathbf{c}_i)^T$$

where \mathbf{c}_i is the class mean defined by $\mathbf{c}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}$ and

$\mathbf{c} = \frac{1}{N} \sum_{i=1}^L \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}$ represents the total mean.

The optimal features are determined by optimizing the Fisher criterion

$J_{LDA} = J_1$ given by

$$J_{LDA} = \text{tr}[(S_w^{LDA})^{-1} S_b^{LDA}].$$

This is equivalent to solving the generalized eigenvalue problem,

$$S_b^{LDA} \mathbf{v}_s = \lambda_s S_w^{LDA} \mathbf{v}_s, \quad s = 1, \dots, d \quad \text{with} \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d,$$

where the extracted eigenvectors form the transformation matrix of LDA.

In other words, the transformation matrix from the original space to the reduced subspace is defined by

$$A = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p].$$

The Fisher criterion J_{LDA} can detect the separability of the transformed training samples, but LDA is a supervised feature extraction. The following section proposes the between- and within-cluster scatter matrices of an unsupervised LDA based on the concept of membership values and cluster means of FCM as a clustering algorithm and an unsupervised feature extraction.

3.2 FLDC Algorithm

The proposed method derives two fuzzy between and within-cluster scatter matrices from the scatter matrices of LDA, and uses them to

formulate FLDC. The fuzzy between-cluster scatter matrix S_b^{FLDA} and the fuzzy within-cluster scatter matrix S_w^{FLDA} are defined as

$$S_b^{FLDA} = \sum_{i=1}^L \frac{\sum_{j=1}^N u_{ij}}{N} (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T$$

and

$$S_w^{FLDA} = \sum_{i=1}^c \sum_{j=1}^N \frac{u_{ij}}{N} (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T,$$

where

$$\mathbf{c}_i = \frac{\sum_{j=1}^N u_{ij}}{\sum_{k=1}^N u_{ik}} \mathbf{x}_j$$

is the class mean, which is the same as FCM, and $\mathbf{c} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_j$ represents the total mean. The following theorem shows that the between- and within-class scatter matrices of LDA are special cases of the proposed S_b^{FLDA} and S_w^{FLDA} , respectively.

Theorem 1: *In the supervised situation, if*

$$u_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in H_i \\ 0 & \text{if } \mathbf{x}_j \notin H_i \end{cases} \text{ for all } 1 \leq i \leq L \text{ and } 1 \leq j \leq N,$$

then, the proposed S_b^{FLDA} and S_w^{FLDA} are the same as S_b^{LDA} and S_w^{LDA} , respectively.

Proof:

Suppose there are N_i samples in H_i for $i = 1, \dots, L$, and $\sum_{k=1}^N u_{ik} = N_i$. Then,

$$\mathbf{c}_i = \sum_{j=1}^N \frac{\mathbf{u}_{ij}}{\sum_{k=1}^N \mathbf{u}_{ik}} \mathbf{x}_j = \sum_{\mathbf{x}_j \in H_i} \frac{1}{N_i} \mathbf{x}_j = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}$$

is the same as the class mean in LDA and the fuzzy between-cluster scatter matrix

$$S_b^{FLDA} = \sum_{i=1}^L \frac{\sum_{j=1}^N \mathbf{u}_{ij}}{N} (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T = \sum_{i=1}^L \frac{N_i}{N} (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T = S_b^{LDA}.$$

The fuzzy within-cluster scatter matrix is then

$$\begin{aligned} S_w^{FLDA} &= \sum_{i=1}^c \sum_{j=1}^N \frac{\mathbf{u}_{ij}}{N} (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T = \sum_{i=1}^c \sum_{\mathbf{x}_j \in H_i} \frac{\mathbf{u}_{ij}}{N} (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T \\ &= \sum_{i=1}^c \sum_{j=1}^{N_i} \frac{1}{N} (\mathbf{x}_j^{(i)} - \mathbf{c}_i)(\mathbf{x}_j^{(i)} - \mathbf{c}_i)^T = S_w^{LDA}. \end{aligned}$$

□

Based on this theorem and the objective function of LDA, the general objective function of FLDC is defined by

$$J_{FLDC}(\mathbf{u}_{ij}) = \text{tr}[(S_w^{FLDA})^{-1} S_b^{FLDA}],$$

including the interaction of S_b^{FLDA} and S_w^{FLDA} . This study considers the interaction of the fuzzy between- and within-cluster scatter matrices in the Fisher criterion. Results for artificial data sets show that FLDC can detect the clusters with the largest between-cluster separability.

To reduce the effects of the cross products of within-class distances and prevent singularity, some regularized techniques [47]-[48] can be applied to the fuzzy within-cluster scatter matrix. In FLDC, the fuzzy within-cluster scatter matrix is regularized by

$$S_{rw}^{FLDA} = rS_w^{FLDA} + (1-r)\text{diag}(S_w^{FLDA})$$

where $\text{diag}(S_w^{FLDA})$ is the diagonal parts of matrix S_w^{FLDA} and $r \in [0,1]$ is a regularization parameter.

The proposed clustering algorithm defines the optimization problem as follows:

$$U_{FLDC} = \arg \max_U J_{FLDC}(u_{ij}) = \arg \max_U [(S_{rw}^{FLDA})^{-1} S_b^{FLDA}]$$

which constrains $\sum_{i=1}^L u_{ij} = 1, j = 1, \dots, N$. Because the optimization problem is nonlinear and non-convex, several popular optimization algorithms [49]-[50] can be applied to solve this problem: “interior-point,” “active-set,” and “trust-region-reflective.” In implementing these algorithms, the “active-set” algorithm has a lower cost time than the other two algorithms, but it is sensitive to the initial value. Hence, the “interior-point” algorithm is used to find the optimizer U_{FLDC} in this study. However, the “interior-point” algorithm has the highest corresponding time cost.

The decision rule, i.e., the defuzzification process, for the sample j is

$$i = \arg \max_k u_{kj}.$$

3.3 Experiments

3.3.1 Experimental Data and Designs

The experiments in this study validate the performance of the proposed FLDC using ten artificial data sets and three real data sets. This section compares the results of several algorithms on artificial and real data sets. These algorithms include the clustering FLDC, K-means (KMS), and K-medoid (KMD), FCM, Gustafson-Kessel (GK), Gath-Geva (GG) [5], possibilistic c-means (PCM) [52], fuzzy-possibilistic c-means (FPCM) [10],

possibilistic fuzzy c-means (PFCM) [11], fuzzy compactness and separation (FCS) [15], FSMM [13], and FMSFA [56] algorithms. The parameters r in FLDC and β in FCS were set to 0.5. The weighting exponents of FCM, GK, GG, and PCM were set to $m \in \{2, 4\}$. The weighting exponents of FPCM and PFCM were set to $m \in \{2, 4\}$ and $\eta \in \{2, 4\}$. The FSMM parameters were set to the default values in [51]. The FMSFA clustering results were the best results within the given set $\{0.5, 1, 1.5\}$ of the model's degrees of fuzziness of the fuzzy membership values.

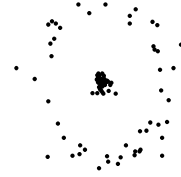
To avoid the influence of initialization, all clustering algorithms were evaluated based on 3 real data sets and 100 randomly generated initial values for each data set. This study calculates and compares the mean, standard deviation, maximum, and minimum accuracy of the 100 clustering accuracy. The accuracy of the clustering is the proportion of correctly clustered data in the data set (i.e., clustering accuracy=(the number of correctly clustered data)/(the number of all samples)).

Fig. 4 shows 10 artificial data sets [53]:“Four-gauss data” (4 clusters), “Easy doughnut data” (2 clusters), “Difficult doughnut data” (2 clusters), “Boat data” (3 clusters), “Noisy lines data” (2 cluster), “Petals data”, (4 clusters), “Saturn data” (2 clusters), “Regular data” (16 clusters), “Half-ring data” (2 cluster), and “Spirals data” (2 clusters). These data sets can be downloaded from [54]. All data sets were created in two dimensions to present challenges in varying degrees. Ten dimensions of uniformly random noise were appended to each of the first three data sets (four gauss, easy doughnut, and difficult doughnut), while the other seven data sets were kept as two-dimensional. The last two data sets were omitted because the clustering results obtained of all clustering algorithms are similar.

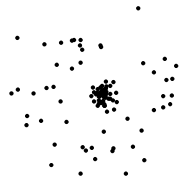
Four gauss (4)



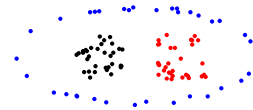
Easy doughnut (2)



Difficult doughnut (2)



Boat (3)



Noisy lines (2)



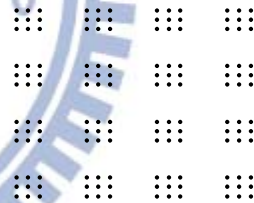
Petals (4)



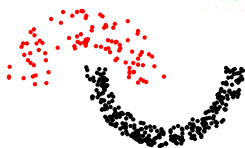
Saturn (2)



Regular (16)



Half rings (2)



Two spirals (2)



Figure 4. Ten artificial data sets [53]-[54] were used in this study. The first three data sets were generated with 10 additional noise features. The number of clusters appears in parentheses.

Table 1 presents the real data sets used in this study: “Wine,” “Iris,” and “Breast Cancer Wisconsin (Diagnostic)” (WDBC). The Wine data set is a collection of data from three classes of wine from various locations in Italy. The Iris data set contains three classes of Iris flowers collected from Hawaii: Iris Setosa, Iris Versicolour, and Iris Virginica. There are two classes, benign and malignant, in the WDBC data set. These data sets are available from the FTP server of the UCI [55] data repository.

Table 1 Descriptions of Three Real Data Sets

Data set	Classes	Number of Samples	Features
Wine	3	178	13
Iris	3	150	4
WDBC	2	569	30

3.3.2 Experimental Results

Figs. 5-12 show the results of clustering on the artificial data sets. The covariance matrices of two density-based methods, GG and FSMM, are near-singular. Hence, the proposed method uses the GG and FSMM with diagonal covariance matrices for the Gaussian distributions (GGD) and the Student’s- t distributions (FSMMD), respectively. The best clustering results from the application of GG and GGD in different data sets were chosen for Fig. 5-12. These figures also show the best results of clustering FSMM and FSMMD. A comparison of Fig. 5-12 reveals the following points:

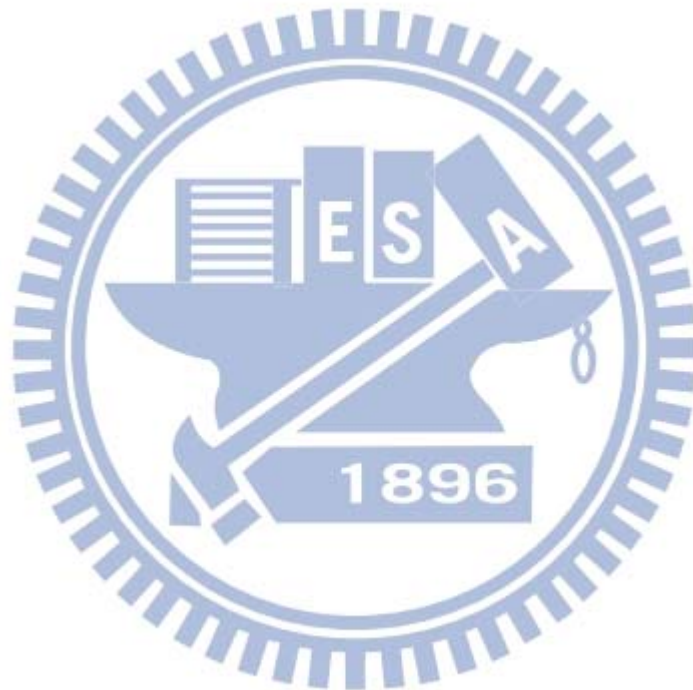
1. The FLDC clustering method significantly outperformed other methods for the normal-like distribution of data (e.g., the four-gauss, easy doughnut, difficult doughnut, boat, and petals data sets) because it

considers the interaction of the between- and within-cluster scatter matrices. For the easy doughnut and difficult doughnut data sets, all algorithms had poor clustering results except FLDC.

2. The FLDC achieved the best performance with regular and noisy lines data sets.
3. KMS, KMD, FCM, FPCM, PFCM, and FCS only performed well on the four gauss and petals data sets.
4. PCM performs well only on the petals and noisy lines data sets.
5. GK employed an adaptive norm that estimates covariance matrices for each cluster. Hence, the GK algorithm can detect clusters with different geometrical shapes. and performed well on the boat and noisy lines data sets. However, its performance was dismal for the four gauss, easy doughnut, and difficult doughnut data sets.
6. Although FLDC performed poorly on the Saturn, half rings, and two spirals data sets, it was able to detect the clusters with the largest between-cluster separability in the Saturn data set. FLDC was unsuitable for the Saturn, half rings, and two spirals data, as these were complex nonlinear problems. The kernel method may be a way to solve these types of data sets.
7. The distribution-based clustering algorithms, including GG, FSMM, and FMSFA, performed poorly on the four gauss, easy doughnut, petals, and regular data sets because the covariance matrices of the density-based methods are near-singular.
8. FSMMD was able to improve the performance of FSMM on the boat and noisy lines data sets.

Table 2 shows the clustering accuracy in real data sets. The highest mean clustering accuracy for each data set (in rows) is shaded. Table 2

shows that the highest mean accuracies among all methods were 0.927, 0.966, and 0.940. All of these results were obtained by performing FLDC. Table 3 shows the accuracy of the three real data sets after applying FMSFA. The maximum accuracies of these data sets were 1, 0.980, and 0.949, respectively. However, it is very sensitive to the initial value. Hence, the highest average accuracies in every column were only 0.945, 0.774, and 0.882. The FSMM is more stable than FMSFA because it uses the results of clustering KMS as the initial value.



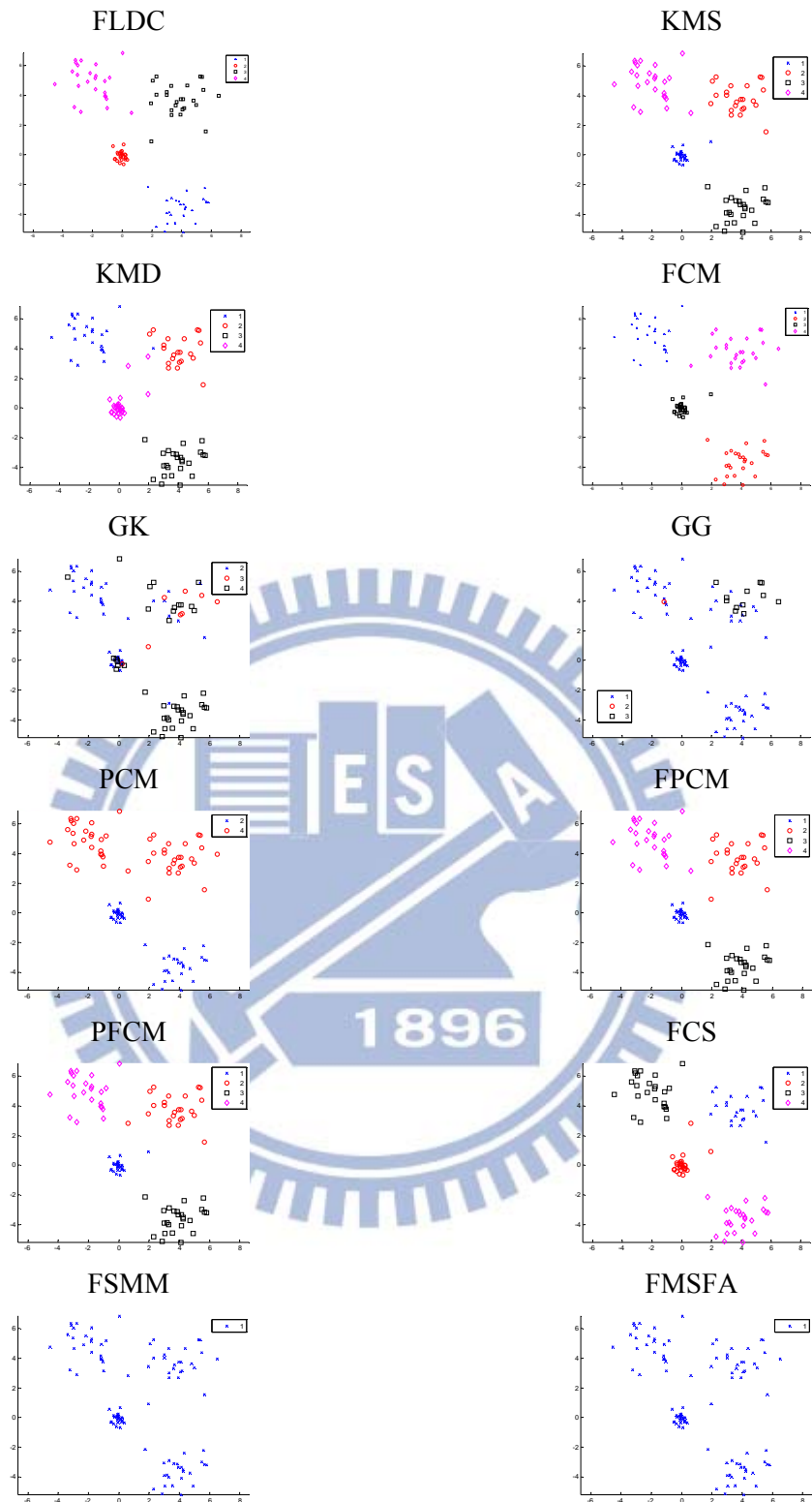


Figure 5. The results of clustering the “Four gauss” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

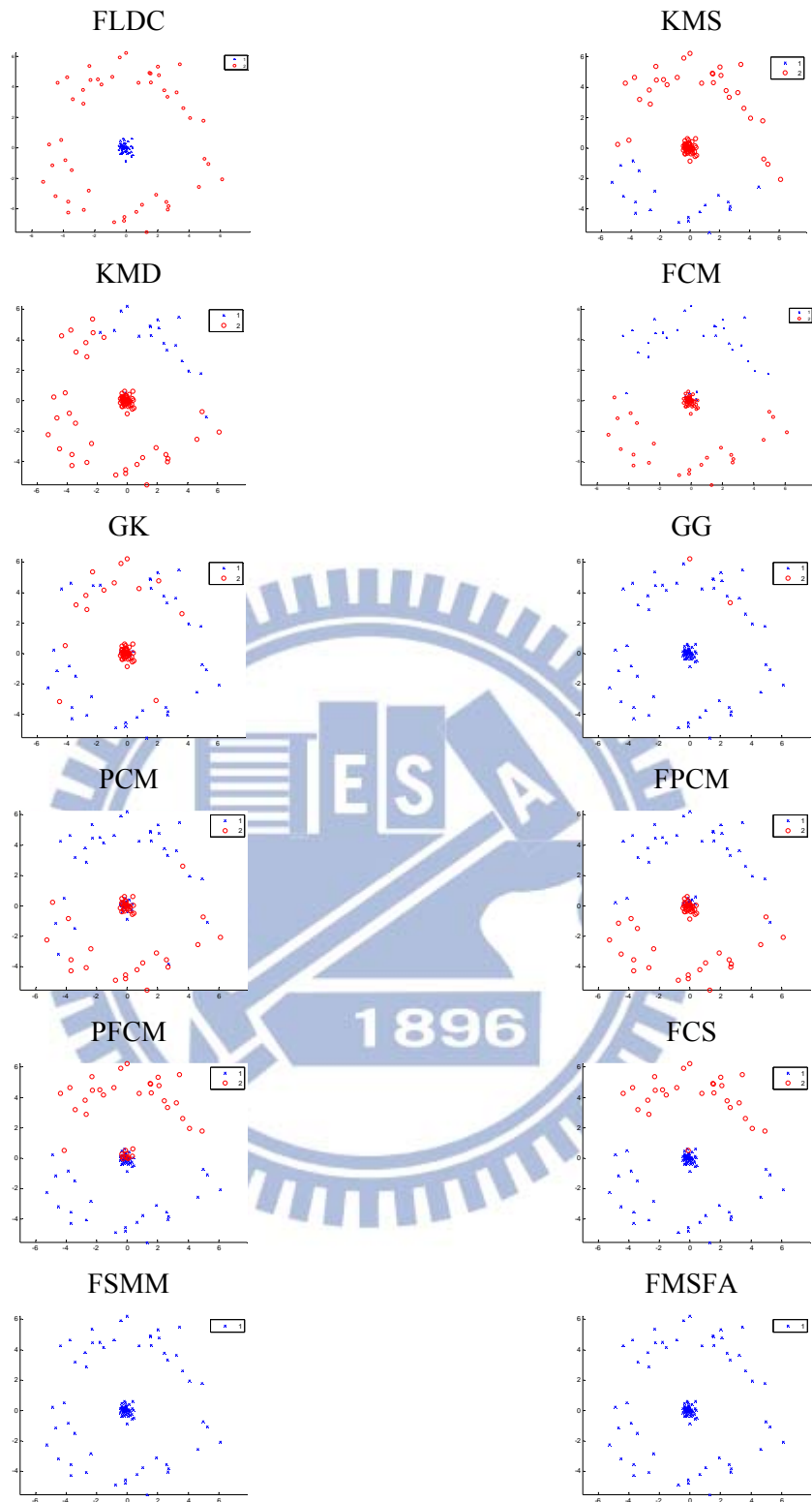


Figure 6. The results of clustering the “Easy doughnut” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

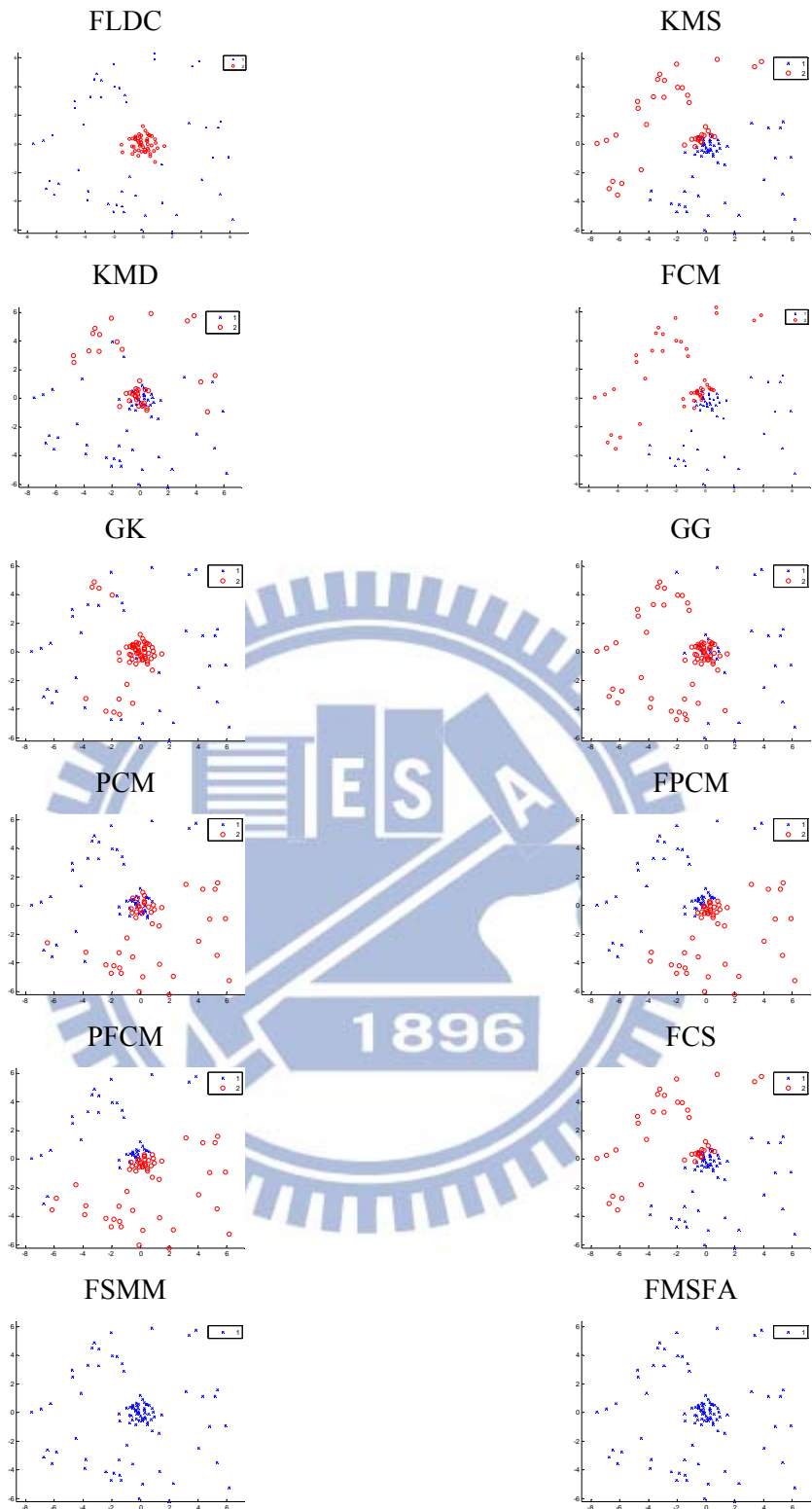


Figure 7. The results of clustering the “Difficult doughnut” data set using twelve clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

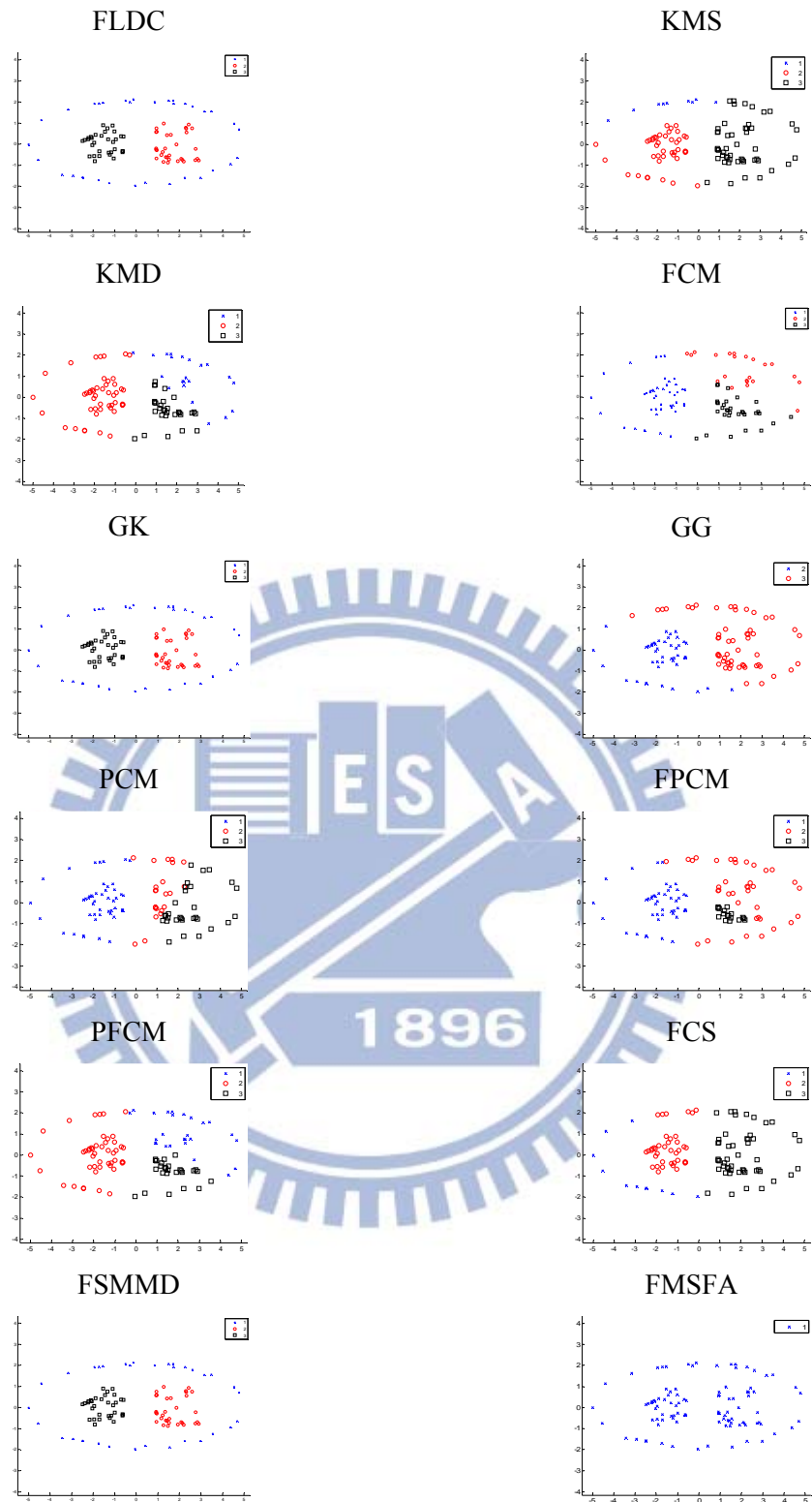


Figure 8. The results of clustering the “Boat” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

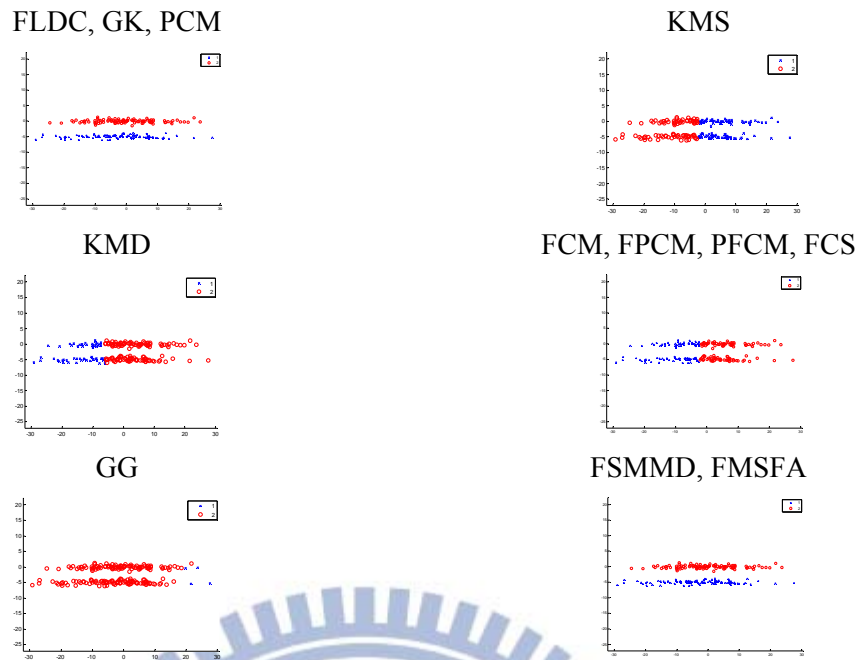


Figure 9. The results of clustering the “Noisy lines” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

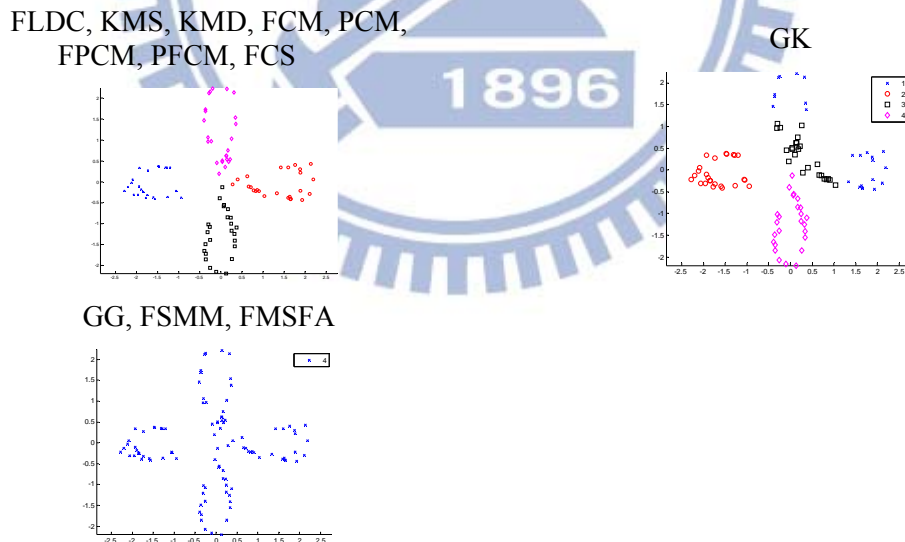


Figure 10. The results of clustering the “Petals” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

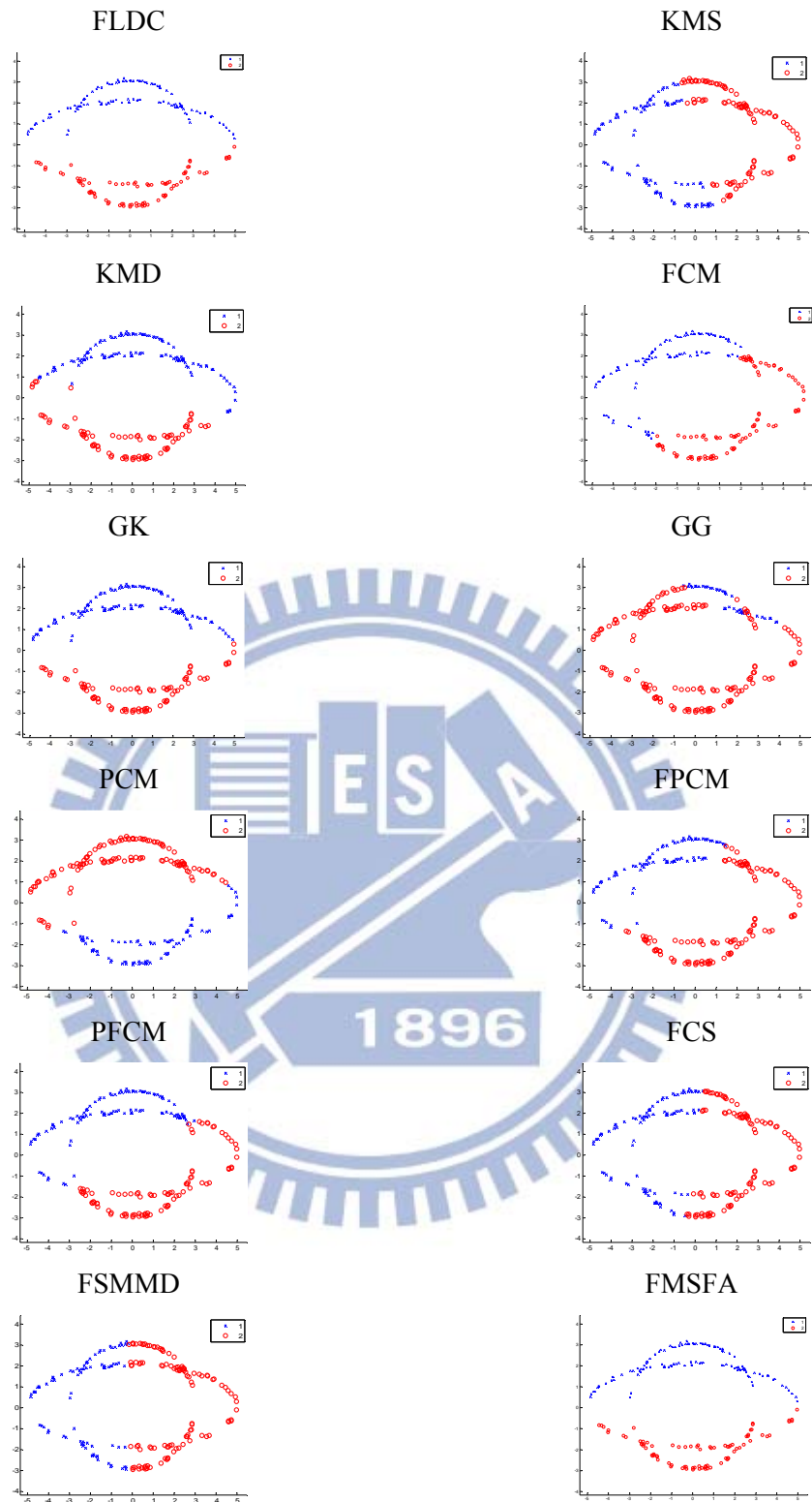


Figure 11. The results of clustering the “Saturn” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

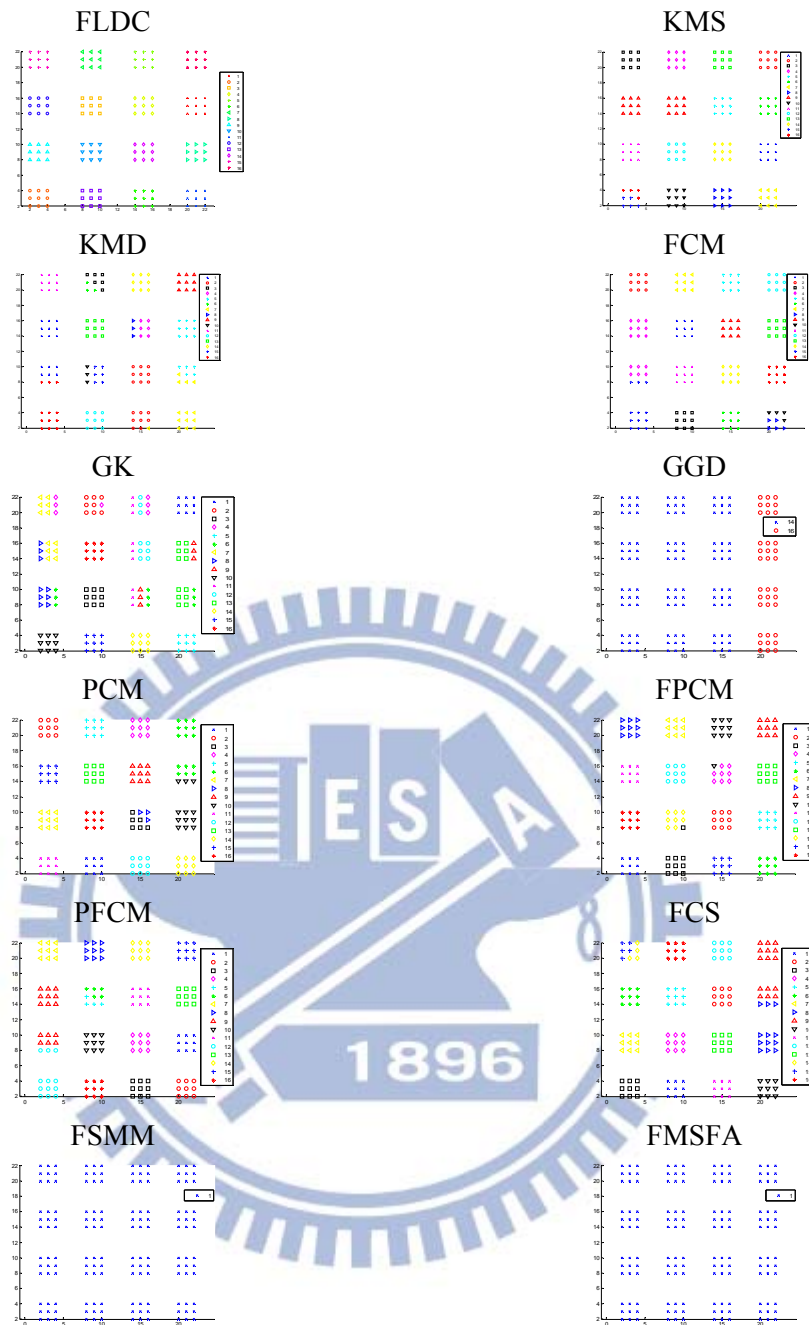


Figure 12. The results of clustering the “Regular” data set using twelve clustering algorithms. The best clustering results from applying GG and GGD were chosen for comparison. This figure also shows the best results of clustering FSMM and FSMMD for comparison.

Table 2 The Mean, Standard Deviation, Maximum, and Minimum Accuracy of Clustering for Three Real Data sets.

	Wine	Iris	WDBC
	mean/std/max/min (mean of cpu time in s.)	mean/std/max/min (mean of cpu time in s.)	mean/std/max/min (mean of cpu time in s.)
FLDC	0.927/0.003/0.949/0.916 (106.628)	0.966/0.001/0.967/0.960 (43.871)	0.940/0.001/0.946/0.938 (2099.937)
KMS	0.677/0.051/0.702/0.567 (0.006)	0.849/0.109/0.893/0.580 (0.009)	0.854/0.000/0.854/0.854 (0.009)
KMD	0.667/0.062/0.708/0.556 (0.004)	0.835/0.141/0.947/0.513 (0.006)	0.851/0.006/0.854/0.837 (0.008)
FCM	0.691/0.000/0.691/0.691 (0.119)	0.907/0.000/0.907/0.907 (0.002)	0.861/0.000/0.861/0.861 (0.108)
GK	0.607/0.000/0.607/0.607 (1.726)	0.900/0.000/0.900/0.900 (0.058)	0.821/0.000/0.821/0.821 (0.254)
GG	0.742/0.000/0.742/0.742 (0.218)	0.733/0.000/0.733/0.733 (0.135)	0.510/0.000/0.510/0.510 (0.113)
PCM	0.697/0.000/0.697/0.697 (0.015)	0.933/0.000/0.933/0.933 (0.014)	0.856/0.000/0.856/0.856 (0.033)
FPCM	0.719/0.000/0.719/0.719 (0.027)	0.907/0.000/0.907/0.907 (0.017)	0.877/0.000/0.877/0.877 (0.036)
PFCM	0.691/0.000/0.691/0.691 (0.060)	0.920/0.000/0.920/0.920 (0.027)	0.861/0.000/0.861/0.861 (0.046)
FCS	0.697/0.000/0.697/0.697 (0.249)	0.893/0.000/0.893/0.893 (0.112)	0.851/0.000/0.851/0.851 (0.321)
FSMM	0.846/0.117/0.899/0.573 (0.583)	0.875/0.170/0.973/0.527 (0.066)	0.935/0.000/0.935/0.935 (0.299)

Table 3 The Mean, Standard Deviation, Maximum, and Minimum Accuracy of Clustering for Three Real Data sets of FMSFA, Where LD Represents the Latent Dimension

	Wine	Iris	WDBC
	mean/std/max/min	mean/std/max/min	mean/std/max/min
FMSFA LD=1	0.898/0.085/0.955/0.854	0.774/0.154/0.980/0.333	0.813/0.005/0.821/0.803
FMSFA LD=2	0.945/0.069/0.966/0.579	0.768/0.127/0.967/0.333	0.882/0.000/0.882/0.882
FMSFA LD=3	0.891/0.143/1.000/0.539	0.704/0.105/0.967/0.333	0.865/0.027/0.949/0.715

4. The Support Vector Machine and Its Spectral-Spatial Classification Schemes

The support vector machine [23] attempts to minimize the upper bound of the generalization error by maximizing the margin between the separating hyperplane and the training data. Hence, SVM is a distribution-free algorithm that can overcome the problem of poor statistical estimation. Many studies [30]-[33] have shown that support vector machines with both spectral and spatial information achieve effective and stable hyperspectral image classification. For example, Tarabalka et al. [31] presented a spectral-spatial classification scheme based on partitional clustering techniques (SVM+EM). A context-sensitive semi-supervised support vector machine (CS⁴VM) [32] uses the context of neighborhood patterns as semi-patterns to solve the problem of noisy training patterns. This chapter reviews the literature on traditional SVM, SVM+EM, and CS⁴VM.

4.1 Support Vector Machine

Let X be a hyperspectral d-dimensional image of size $I \times J$ pixels. Assume that a set of training data set

$$D = \{ x_i \mid x_i \in X \subset R^d, i = 1, 2, \dots, n \}$$

is available and $\{ y_i \in \{+1, -1\} \}_{i=1}^n$ is the corresponding label set. SVM tries to find a separating hyperplane in the feature space, a Hilbert space H , for a binary classification problem [23]. The soft-margin SVM algorithm is based on the following constrained minimization optimal problem:

$$\min_{\mathbf{w}, \xi} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i \quad (4.1)$$

$$\text{subject to } y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, n$$

where \mathbf{w} is a vector normal to the hyperplane, b is a constant such that $b/\|\mathbf{w}\|$ represents the distance between hyperplane from the origin, $\phi: R^d \rightarrow H$ is a nonlinear mapping function, ξ_i 's are slack variables to control the training errors, $\xi = [\xi_1, \dots, \xi_n]^T$, and $C \in R^+$ is a penalty parameter for tuning the generalization capability. Trying to solve this optimal problem with inequality constraints is generally a difficult task. However, the original optimal problem has an equivalent dual representation using the Lagrange optimization method. The corresponding dual Lagrange function is defined as

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n \end{aligned} \quad (4.2)$$

where the artificial variables, α_i , are Lagrange multipliers, and $\alpha = [\alpha_1, \dots, \alpha_n]^T$.

The kernel trick uses a kernel function $\kappa: R^d \times R^d \rightarrow R$ to implicitly map the data from the original space R^d to H without knowing the feature mapping ϕ . The inner product of samples in the feature space can be computed directly from the original data items using a kernel function. This is because a kernel function κ satisfies Mercer's theorem [34]. In other words, there is a feature map ϕ into a Hilbert space H such that $\kappa(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z})$, where $\mathbf{x}, \mathbf{z} \in R^d$, if and only if κ is a symmetric

function for which the matrices $K = [\kappa(x_i, x_j)]_{1 \leq i, j \leq n}$ formed by restriction to any finite subset $\{x_1, \dots, x_n\}$ of the space R^d are positive semi-definite. Hence, the kernel trick makes it possible to rewrite Eq. (4.2) as the following Eq. (4.3). Since, for a kernel function, the corresponding kernel matrix is positive semi-definite for all training sets, this in turn means that the optimization problem of (4.3) is always convex [34].

$$\max_{\alpha} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \quad (4.3)$$

$$\text{subject to } \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n$$

After determining the values of the α_i 's, the decision function for an unlabeled pattern x is defined as

$$f_{\text{SVM}}(x) = \sum_{i=1}^n y_i \alpha_i \kappa(x_i, x) + b,$$

where b is chosen so that $y_j (\sum_{i=1}^n y_i \alpha_i \kappa(x_i, x_j) + b) = 1$ for any x_j with $0 < \alpha_j < C$, and a corresponding forecasting label is $\text{sgn}(f_{\text{SVM}}(x))$.

4.2 Spectral-Spatial Classification Scheme Based on Partitional Clustering Techniques

Fig. 13 shows a flowchart of the spectral–spatial classification scheme based on partitional clustering techniques (SVM+EM) [31]. The majority vote rule is used to determine the final decision in a partition.

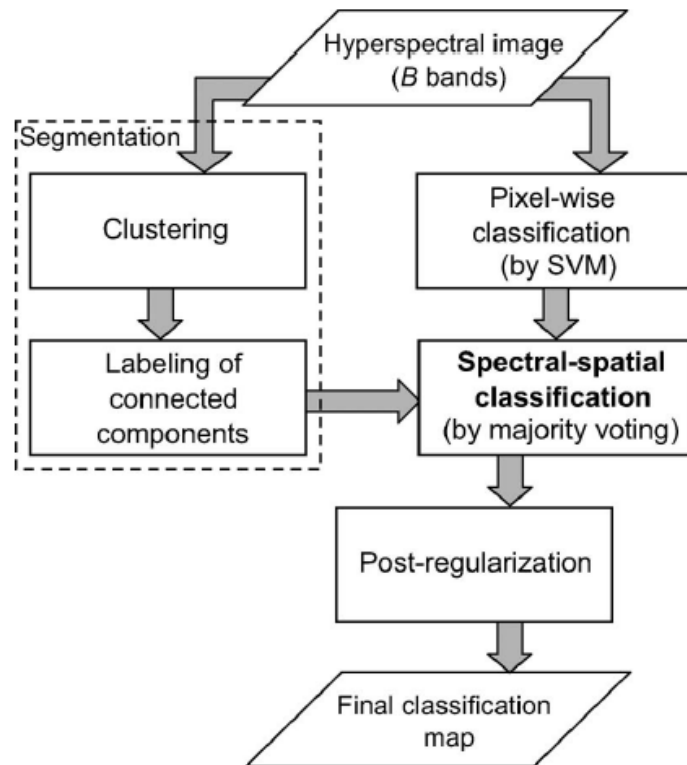


Figure 13. Flowchart of the SVM+EM [31].

SVM+EM combines the unsupervised segmentation technique and supervised pixel wise classification results, and consists of the following steps (Fig. 13 and 14).

1. *Segmentation*: The expectation maximization (EM) clustering algorithm segments a hyperspectral image into homogeneous regions.
2. *Pixel wise classification*: An SVM classifier with the Gaussian radial basis function (RBF) kernel is performed independently of the

segmentation procedure. The SVM parameters are determined by k -fold cross validation.

3. *Spectral–spatial classification*: The majority rule is used for every region in the segmentation map. All samples in the same region are assigned to the most frequent class within this region.
4. *Spatial postregularization (PR) step*: Finally, the spatial PR of the classification map reduces the noise in the classification map.

SVM+EM segments an image into homogeneous regions and combines the results of these regions using pixel-wise SVM classification. The spatial post regularization (PR) of the classification map reduces the noise. This approach is particularly suitable for classifying images with large spatial structures, when spectral responses of different classes are dissimilar, and the classes contain a similar number of pixels. If the spectral responses are not significantly different, this approach may result in misclassification [31].

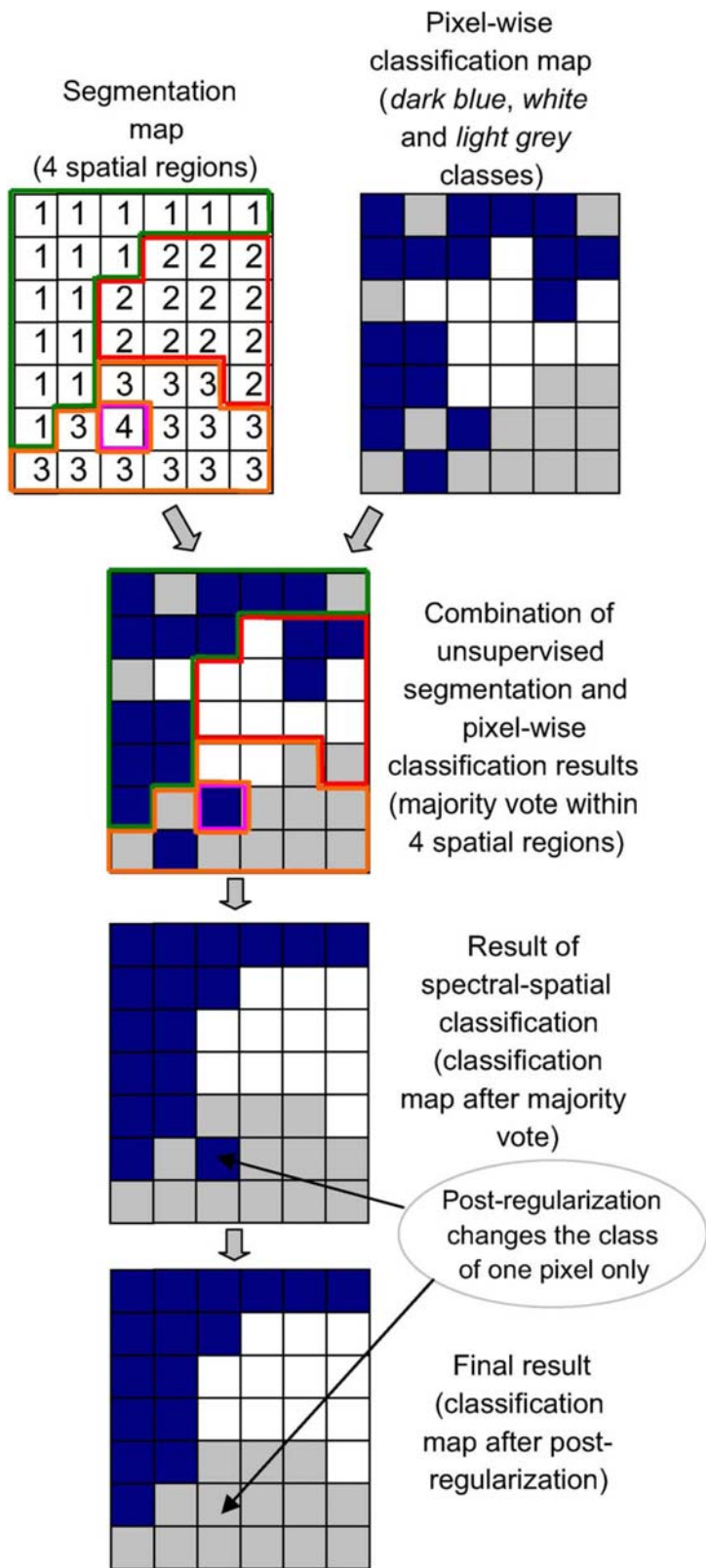


Figure 14. Example of SVM+EM classification [31].

4.3 Context-Sensitive Semi-supervised SVM

The context-sensitive semi-supervised SVM (CS⁴VM) classifier improves robustness to possible mislabeled training patterns by exploiting the contextual information of the pixels belonging to the neighborhood system of each training sample in the learning phase [32]. Let $\partial x_i^o = \{\bar{x}_{ij} | j=1, \dots, M\}$ represent a neighborhood system of the pixel x_i in the original space, where M is 4 or 8 to indicate that ∂x_i^o is a first-order or second-order neighborhood system, respectively (Fig. 15). After performing the standard SVM, CS⁴VM can obtain the semi-labels of ∂x_i , which is equal to ∂x_i^o and denote them as $\{y_{ij}\}_{j=1}^M$ (i.e., $y_{ij} = \text{sgn}(f_{\text{SVM}}(\bar{x}_{ij}))$, $i=1, \dots, n$, $j=1, \dots, M$). The cost function of CS⁴VM for the learning of the classifier is

$$\begin{aligned}
 \min_{\mathbf{w}, \xi, \psi} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \sum_{j=1}^M \kappa_{ij} \psi_{ij} \\
 \text{subject to} \quad & y_i (\mathbf{w}^T \phi(x_i) + b) \geq 1 - \xi_i, \\
 & y_{ij} (\mathbf{w}^T \phi(\bar{x}_{ij}) + b) \geq 1 - \psi_{ij} \\
 & \xi_i, \psi_{ij} \geq 0, \quad i=1, 2, \dots, n, \quad j=1, 2, \dots, M
 \end{aligned} \tag{4.4}$$

where ψ_{ij} 's are context slack variables and $\kappa_{ij} \in R^+ \cup \{0\}$ are parameters that make it possible to weight the importance of context patterns (Fig. 16). The aim of the cost function of CS⁴VM is to regularize the learning process with respect to the behavior of the context patterns in the neighborhood of the training pattern under consideration. This term helps balance the contribution of possibly mislabeled training samples according to the semi-labeled pixels of the neighborhood [32].

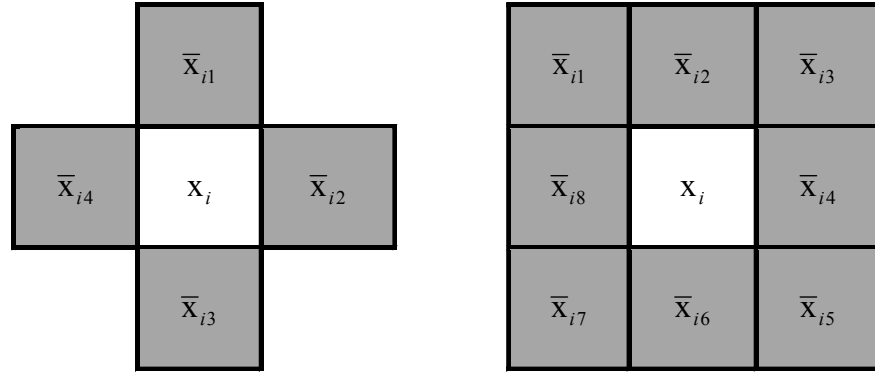


Figure 15. The left and right images represent the first-order and second-order neighborhood systems in the original space, respectively.

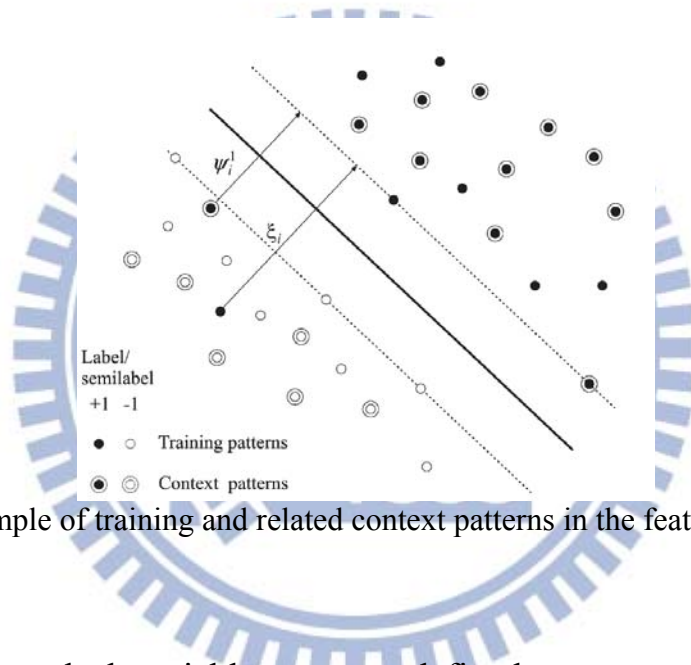


Figure 16. Example of training and related context patterns in the feature space [32].

The context slack variables ψ_{ij} are defined as

$$\psi_{ij} = \max\{0, 1 - y_{ij}(\mathbf{w}^T \phi(\bar{x}_{ij}) + b)\}, \quad \forall i = 1, 2, \dots, n, \quad \forall j = 1, 2, \dots, M.$$

The parameters κ_{ij} weight the context patterns \bar{x}_{ij} depending on the agreement of their semi-labels y_{ij} with that of the related label y_i of the training sample x_i . The hypothesis at the basis of the weighting system of the context patterns is that the pixels in the same neighborhood system are likely to be associated with the same information class (i.e., the labels of

the pixels are characterized by high spatial correlation). In particular, κ_{ij} 's are defined as

$$\kappa_{ij} = \begin{cases} \kappa_1 & \text{if } y_i = y_{ij} \\ \kappa_2 & \text{if } y_i \neq y_{ij} \end{cases},$$

where κ_1 and κ_2 are chosen by the user to define the importance of the context patterns. It is very important to define the ratios $C/\kappa_i, i=1, 2$, which tune the weight of context patterns w.r.t. the patterns of the original training set. The selection of κ_1 and κ_2 can be simplified by fixing a priori the ratio $\kappa_1/\kappa_2 = K$. This focuses attention only on κ_1 or on the ratio C/κ_1 [32].

CS⁴VM uses the context of neighborhood patterns as semi-patterns to solve the problem of noisy training patterns. In this case, noisy training patterns are mislabeled patterns that introduce distorted information to the classifier [32]. However, CS⁴VM is a semi-learning approach in which the computational cost increases as the number of semi-samples increases.

5. Spatial-Contextual Support Vector Machines

The two sections of this chapter introduce two kinds of spatial-contextual support vector machines with different neighborhood systems: the original space (SCSVM) and the feature space (SCSVMF) [39]. The learning process of the proposed SCSVM classification system includes three steps: i) learning the standard SVM to classify the image, ii) learning SCSVM/SCSVMF with both spectral and spatial-contextual information, and iii) repeating (ii) to update the unlabeled patterns until convergence.

5.1 A Spatial-contextual Support Vector Machine in the Original Space

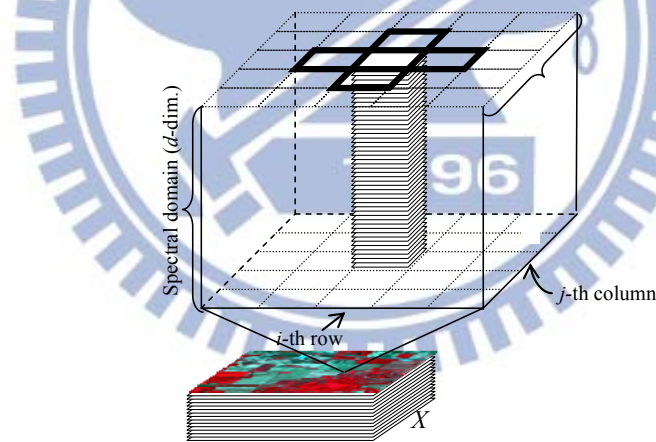


Figure 17. The pixels enclosed by bold lines represent the first-order neighborhood system used in SCSVM.

In SCSVM, spatial information exploits the semi-labels for the pixels belonging to the neighborhood system in the original space (Fig. 17) of each sample from the preceding discriminated process of standard SVM to overcome similar spectral properties. SCSVM can achieve good generalization, especially for pixels with similar spectral attributes but

located in different regions. This approach decreases speckle-like errors and significantly improves classification performance. Let $\partial x_i^o = \{\bar{x}_{ij} \mid j = 1, \dots, M\}$ represent a neighborhood system of the pixel x_i in the original space, where M is 4 or 8 to represent that ∂x_i^o is a first-order or second-order neighborhood system, respectively.

After performing the standard SVM, SCSVM can obtain the semi-labels of ∂x_i , which is equal to ∂x_i^o , and denote them as $\{y_{ij}\}_{j=1}^M$ (i.e., $y_{ij} = \text{sgn}(f_{\text{SVM}}(\bar{x}_{ij}))$, $i = 1, \dots, n$, $j = 1, \dots, M$). The constrained minimization problem associated with SCSVMs accounts for the semi-labels of the whole image, and is defined as follows:

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b + \gamma (m^+(\mathbf{x}_i) - m^-(\mathbf{x}_i))) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, 2, \dots, n \end{aligned} \tag{5.1}$$

where $\gamma \in [0, \infty)$ is a nonnegative parameter that controls the effects of spatial-contextual information. $m^+(\mathbf{x}_i)$ and $m^-(\mathbf{x}_i)$ represent the number of pixels in the neighbor system ∂x_i that belongs to class +1 and class -1, respectively. Fig. 18 illustrates the spatial-contextual information of the pattern x_i with the second-order neighborhood system $\partial x_i = \partial x_i^o$ employed in the spatial domain.

The SCSVM cost function does not require modification, and maintains the property of convex property. Because the objective function of the minimization problem of the SCSVM only contains training samples, and no semi-label samples, the decision hyperplane is not influenced by samples with similar spectra. The computational costs of each iteration in SCSVM are also similar to that of SVM.

$y_{i1}^s = +1$	$y_{i2}^s = -1$	$y_{i3}^s = -1$
$y_{i8}^s = +1$	y_i	$y_{i4}^s = +1$
$y_{i7}^s = +1$	$y_{i6}^s = +1$	$y_{i5}^s = +1$

$m^+(x_i) = 6$
 $m^-(x_i) = 2$

Figure 18. An example of the spatial-contextual information with the second-order neighborhood system of pattern x_i in the original space.

According to Lagrange's theorem, the corresponding dual problem is as follows:

$$\begin{aligned}
 \max_{\alpha} \quad & \sum_{i=1}^n (1 - y_i \gamma (m^+(x_i) - m^-(x_i))) \alpha_i \\
 & - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \\
 \text{subject to} \quad & \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n
 \end{aligned} \tag{5.2}$$

when $\alpha_i, i = 1, \dots, n$ are determined, the decision function for an unlabeled pattern x is defined as

$$f_{\text{SCSVM}}(x) = \sum_{i=1}^n y_i \alpha_i \kappa(x_i, x) + b + \gamma (m^+(x) - m^-(x)). \tag{5.3}$$

Any generic pattern belonging to the investigated image can then be classified according to

$$\text{sgn}(f_{\text{SCSVM}}(x)).$$

If some training patterns appear in the margin, they may produce similar spectral properties. Hence, these patterns may be noisy patterns in standard

SVM learning. To overcome this problem, the constraints and the decision function of SCSVM include spatial terms. If

$$m^+(x_j) - m^-(x_j) > 0 \quad \text{and} \quad m^+(x_j) - m^-(x_j) < 0,$$

then $f_{\text{SCSVM}}(x_j) > f_{\text{SVM}}(x_j)$ and $f_{\text{SCSVM}}(x_j) < f_{\text{SVM}}(x_j)$, respectively. This means that if the semi-labels of most patterns in the neighborhood system ∂x_j are +1, then the signed distance from x_j to the decision hyperplane of SCSVMs will tend to be positive. If the semi-labels of most patterns in the neighborhood system ∂x_j are -1, then the signed distance from x_j to the decision hyperplane of SCSVM will tend to be negative. The parameter γ controls the effect of the spatial-contextual information (i.e., the term, $m^+(x_j) - m^-(x_j)$). If γ is set to 0, then SCSVM degenerates to the standard SVM. When γ increases, the effect of neighborhood points (spatial information) increases. If γ approaches ∞ , then the semi-label of x_j is determined by the sign of $m^+(x_j) - m^-(x_j)$ (i.e., the spatial-contextual information).

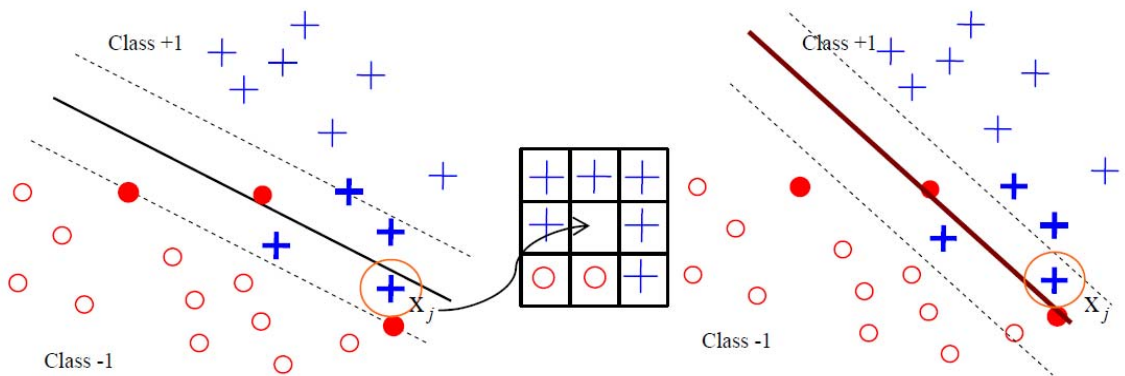


Figure 19. The left panel shows the decision boundary (solid black line) obtained by SVM. The center panel shows the semi-labels of the patterns in the second-order neighborhood system of x_j . The right panel shows the decision boundary (solid red line) obtained of SCSVM.

Fig. 19 shows the effects of applying SCSVM. The left panel shows the decision boundary (solid black line) obtained of standard SVM. The training sample x_j with $y_j = +1$ is in the opposite area (class -1) but in the area between margins. After performing standard SVM, $m^+(x_j) - m^-(x_j) = 4 > 0$. The spatial-contextual information in the center panel of Fig. 20 shows that the training sample x_j should be in the area in which sample labels are 1. If $f_{SCSVM}(x_j) > f_{SVM}(x_j)$, then x_j would be in the expected area (class +1), as shown in the right panel of Fig. 19.

As mentioned, SCSVM depends on the spectral information and the spatial-contextual information, which is based on the neighborhood system in the original system. Hence, the problem of similar spectral properties can be solved of SCSVM. SCSVM applies the spatial-contextual information to emphasize the effects of this pattern on the learning phase.

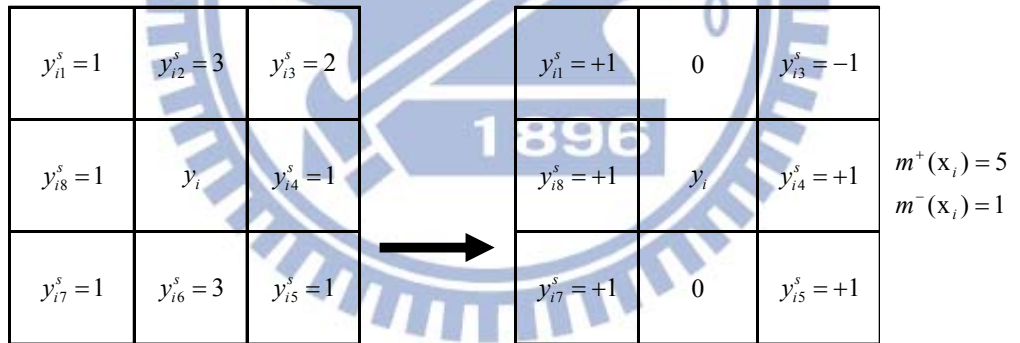


Figure 20. A multiclass case of the spatial contextual information defined by the OAO strategy (class 1 versus class 2) for pattern x_i in the neighborhood system ∂x_i^O . The labels of class 1 and class 2 are defined as +1 and -1, respectively.

To address the multiclass classification problem, the following paragraphs describe the two types of SCSVM for multiclass strategies: the one-against-one (OAO) strategy [60]-[62] and the one-against-all (OAA) strategy [62]. The OAO strategy separates each pair of classes. Thus, for a

classification problem with L classes, $L(L-1)/2$ SCSVMs are trained to distinguish the samples of one class from the samples of another class. The classification results of an unlabeled pattern are based on the maximum vote, where each SCSVM votes for one class. When an SCSVM is trained by two classes of training data, it ignores the spatial-contextual information of other classes to avoid misjudgments in training process.

Fig. 20 shows the OAO strategy for computing $m^+(x_i)$ and $m^-(x_i)$ in the neighborhood system ∂x_i^o of x_i . Suppose there are 3 classes and SCSVM is trained by the training samples in class 1 and class 2. Thus, all semi-labels equal to 3 will be omitted. Since $y_{i2}^s = 3$ and $y_{i6}^s = 3$, these spatial-contextual information are ignored and, hence, $m^+(x_i) = 5$ and $m^-(x_i) = 1$.

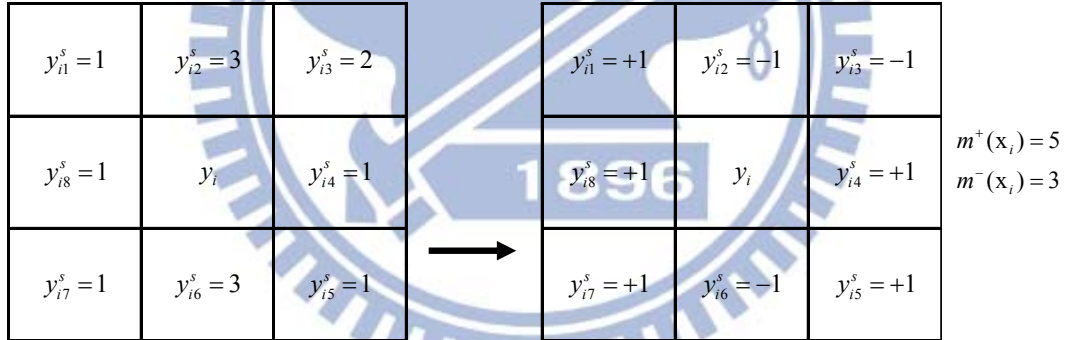


Figure 21. A multiclass case of the spatial contextual information defined by the OAA strategy (class 1 versus all others) for pattern x_i in the neighborhood system ∂x_i^o . The label of class 1 is defined as +1 and the labels of the remaining classes (class 2 and class 3) are defined as -1.

The one-against-all (OAA) multiclass strategy trains L SCSVMs, one per class, using members of all other classes as negative examples if there are L classes. Fig. 21 shows the OAA strategy for computing $m^+(x_i)$ and $m^-(x_i)$ in the neighborhood system of the pattern x_i . When the k -th

OAA SCSVM is trained, the class k is set as the positive class and other classes are all set as negative class. Fig. 21 shows $m^+(x_i) = 5$ and $m^-(x_i) = 3$ for the example by considering the neighborhood system ∂x_i .

5.2 A Spatial-Contextual Support Vector Machine in the Feature Space

The SCSVMF approach uses the same concept as SCSVM except that the neighborhood system $\partial x_i = \partial x_i^F$, which contains M nearest neighbors in the feature space:

$$\partial x_i^F = \{ \bar{x}_{ij} \mid \| \phi(x_i) - \phi(x_{ij}) \| \leq \| \phi(x_i) - \phi(z) \|, \forall z \in X, z \neq x_i, z \neq x_{ij}, j = 1, \dots, M \}.$$

Similar spectral properties cannot be solved efficiently of SCSVMF because the nearest neighbors in the feature space are used in the neighborhood system ∂x_i^F . Hence, being neighborhoods of a given point in feature space is caused by the similar spectra. These neighborhood points may not have geographic relationship. The classification accuracy may also decrease when the neighborhood points in feature space are from different classes.

When SCSVMF is trained by the training samples in class k and class s , the OAO strategy ignores the semi-labels of samples in ∂x_i^F that are not equal to k and s in the multiclass classification problem. The OAA multiclass strategy sets the semi-labels of samples in ∂x_i^F that belong to class k as the positive class and other semi-labels of samples in ∂x_i^F as negative classes when training the k -th OAA SCSVM. Similarly, if γ is set to 0, then SCSVMF degenerates to the standard SVM.

5.3 Classification System of SCSVM and SCSVMF

Based on these descriptions and definitions, Fig. 22 illustrates the proposed SCSVM and SCSVMF classification systems.

- Step 1: Obtain the classification image with semi-labels from the standard SVM.
- Step 2: Acquire the spatial-contextual information for each training pattern with OAO or OAA multiclass architecture from the preceding classification result.
- Step 3: Train the proposed SCSVM (SCSVMF) with the spatial-contextual information from Step 2, and get another classification image with the semi-labels obtained from SCSVM (SCSVMF).
- Step 4: Repeat Steps 2 and 3 if an iteration is requested. The iteration may terminate when the difference of semi-labels in this iteration step and the previous iteration step is smaller than a certain tolerance value.
- Step 5: Perform the spatial post regularization (PR) of the classification map for SCSVM [31]. This PR step attempts to reduce the noise in the classification map after the majority vote procedure.

Fig. 22 shows the framework of the SCSVM (SCSVMF) algorithm.

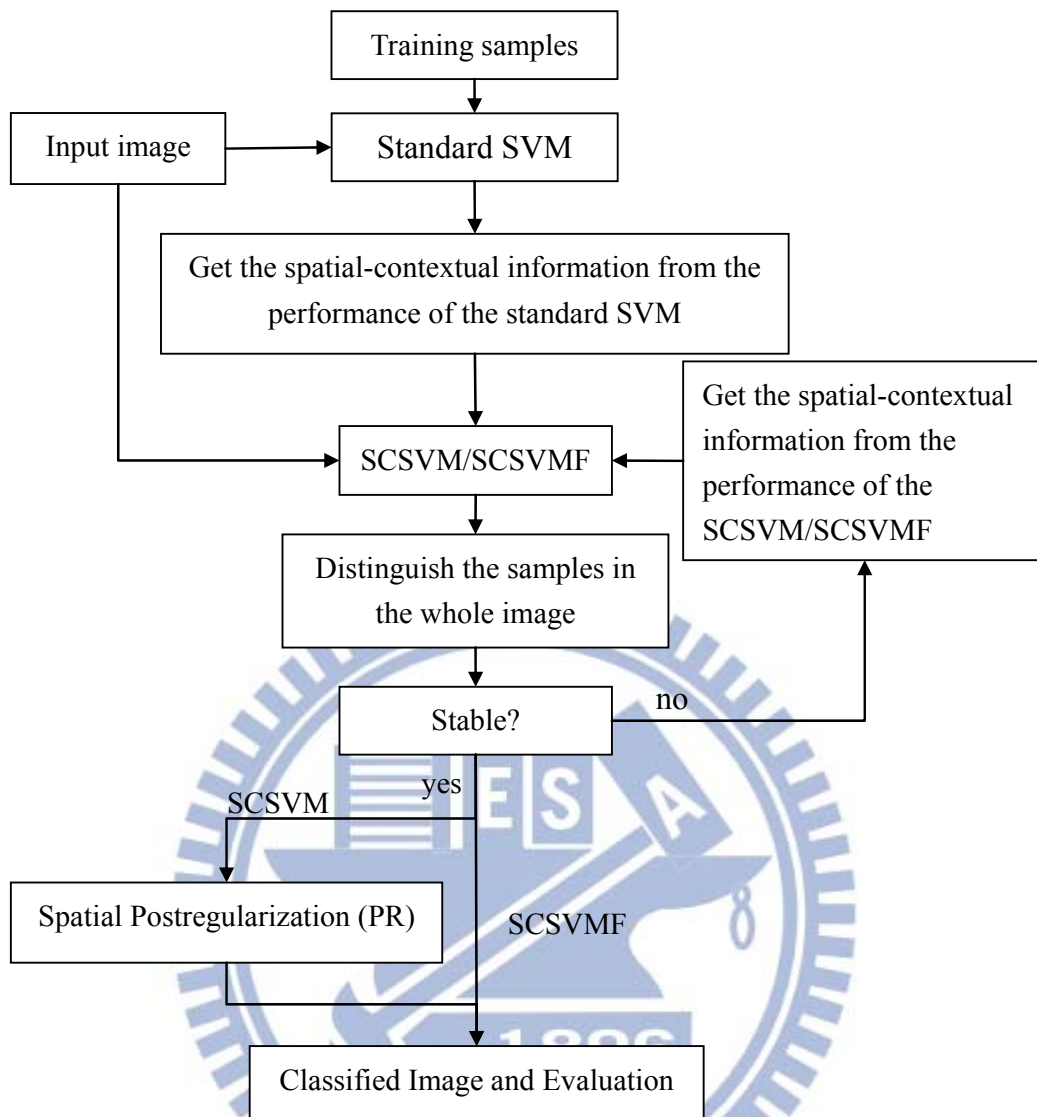


Figure 22. SCSVM and SCSVMF classification systems.

5.4 Experiments

5.4.1 Experimental Data and Designs

The experiments in this study use two real data sets to evaluate the classification performance of the proposed SCSVM and SCSVMF: the Indian Pine Site (IPS), a mixed forest/agricultural site in Indiana [22], and a hyperspectral image of the Washington D.C. Mall [22] as an urban site.

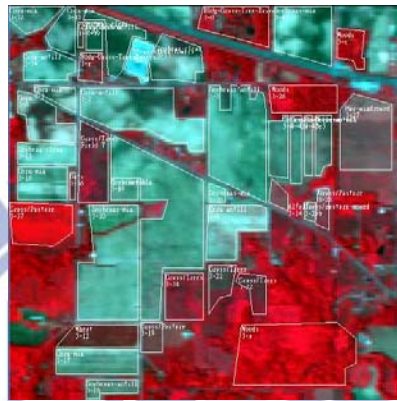


Figure 23. A portion of the Indian pine site image measuring 145×145 pixels.



Figure 24. The ground truth of the Indian pine site data set.

The IPS data set was gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). This data set was obtained from an aircraft operated by the NASA/Jet Propulsion Laboratory flying at an altitude of 65000 ft. Each images measures 145×145 pixels, with 220 spectral bands measuring approximately 20 m across the ground. Figs. 23 and 24 show the grayscale IR image and ground truth of IPS, respectively. The original ground-truth image contains 16 different land-cover classes. This study uses sixteen categories: Alfalfa (class 1), Corn-no till (class 2), Corn-min till (class 3), Corn (class 4), Hay-windowed (class 5), Grass/trees (class 6), Grass/pasture-mowed (class 7), Grass/pasture (class 8), Oats (class 9), Soybeans-no till (class 10), Soybeans-min till (class 11), Soybeans-clean till (class 12), Wheat (class 13), Woods (class 14), Bldg-Grass-Tree-Drives (class 15), and Stone-steel towers (class 16). Table 4 lists the number of pixels of each class.

Table 4 Sixteen Categories and Corresponding Number of Pixels in the Indian Pine Site Image

No.	Category	#(pixels)	No.	Category	#(pixels)
1	Alfalfa	46	9	Oats	20
2	Corn-no till	1428	10	Soybeans-no till	972
3	Corn-min till	830	11	Soybeans-min till	2455
4	Corn	237	12	Soybeans-clean till	593
5	Hay-windowed	483	13	Wheat	205
6	Grass/trees	730	14	Woods	1265
7	Grass/pasture-mowed	28	15	Bldg-Grass-Tree-Drives	386
8	Grass/pasture	478	16	Stone-steel towers	93

This experiment randomly chose ten percent of the samples for each class from the IPS reference data as training samples, following the method in [31]. The samples in the whole image served as the testing set to evaluate the performance of the proposed algorithm.

The second data set, the Washington D.C. Mall, was obtained in a Hyperspectral Digital Imagery Collection Experiment (HYDICE). Images were acquired from an airborne hyperspectral data flightline over Washington D.C. In total, 210 bands were collected in the 0.4-2.4 μm region of the visible and infrared spectrum. Some water absorption channels were discarded, resulting in 191 channels. This data set is available in the student CD-ROM of [22]. The second experiment in this study used 7 classes: grass (class 1), tree (class 2), roof (class 3), water (class 4), road (class 5), trail (class 6), and shadow (class 7). Fig. 25 shows the grayscale IR image of a portion of the image and the seven corresponding categories.

No.	Category
1	Grass
2	Tree
3	Roof
4	Water
5	Road
6	Trail
7	Shadow




Figure 25. The false-color IR image of a portion of Washington D.C. Mall image measuring 205 307 pixels. There are seven categories: grass, tree, roof, water, road, trail, and shadow.

This study uses three distinct subsets, $N_i=20 < N < d$ (case 1), $N_i=40 < d < N$ (case 2), and $d < N_i=300 < N$ (case 3), to investigate the influence of training sample size on the dimensionality of the Washington D.C. hyperspectral image data set. In case 1, $N_i = 20 < N=180 < d = 191$ is an ill-posed classification situation, which means data dimensionality exceeds the number of independent training samples in every class. In case 2, $N_i = 40 < d = 191 < N = 360$ is a poorly posed classification situation, which means that data dimensionality is greater than or comparable to the number of (independent) per-class representative training samples, but smaller than

the total number of representative samples. In case 3, there are enough independent training samples. MultiSpec [22] was used to randomly select training and testing samples (100 testing samples per class) in all experiments [63]-[65].

This study compares the classification performance of the proposed SCSVM and SCSVMF with OAO and OAA multiclass strategies and other reference classification algorithms: ML classifier [2], ML_MRF classifier [20], k -NN classifier [2], standard SVM with OAO and OAA multiclass strategies, CS⁴VM (which is based on the OAA multiclass strategy) [32], and SVM+EM [31]. This experiment also compares the classification performance of SVM+EM and SCSVM with the PR step using a 3×3 mask and without the PR step. The SVM-based classifiers, including SVM, CS⁴VM, and SCSVM, employ the RBF kernel (i.e., the Gaussian Radial Basis Function kernel). Both the IPS and the Washington D.C. Mall hyperspectral data sets were normalized to the range [0, 1]. A grid search with k -fold cross validation was used to find the proper $2\sigma^2$ within a range $[10^{-2}, 10]$ for the RBF kernel (as suggested by [32]) and parameter C within a given set $\{0.1, 1, 10, 20, 60, 100, 160, 200, 1000\}$. For CS⁴VM, the value of κ_1/κ_2 was set to 2 and $C/\kappa_1 \in \{2, 4, 6, 8, 10, 12, 14\}$ following [32]. Because the semi-samples were used to train CS⁴VM, only a first-order neighborhood system ∂x_i^o was considered for the context patterns to avoid spending too much time training CS⁴VM. For SCSVM, only the decision function and constraints contain the spatial-contextual information of the neighborhood system. Thus, the SCSVM training time increases a little for each iteration. The size of the neighborhood system M was set to 4 and 8 in SCSVM for comparison. The term γ was set to 0.05, 0.1, 0.3, 0.5, 1, 10, 100, 500, 1000, and 10000 to determine its influence on spectral and spatial information.

The Gaussian function was adopted as the likelihood function of the Bayesian decision rule for the ML classifier and ML_MRF classifier [20]. Several trials were carried out for the k -NN classifier, varying the value of k from 1 to 20 to identify the value that maximizes the accuracy. For simplicity, the model selection for the k -NN classifier was based on the accuracy of the testing data set.

This study employs the following measures of classification accuracy to investigate classifier performance: 1) overall classification accuracy (the percentage of correctly classified samples for all classes); 2) overall kappa coefficient (the percentage of the kappa coefficient for all classes); and 3) average accuracy (the average percentage of correctly classified samples for each class). Because the amount of testing data is the same for every class (i.e., $N_i=100$) in the Washington D.C. hyperspectral image data set. In this case, the overall classification accuracy and the average accuracy are identical. In the IPS data set, the overall classification accuracy and the average accuracy are not identical because of the unequal testing sample sizes between classes.

5.4.2 Experimental Results

This study compares the multiclass-classification performance of an ML classifier, ML_MRF classifier, k -NN classifier, SVM, CS⁴VM, SVM+EM, and SCSVM. The following section presents the experimental results for the IPS data set and the Washington D.C. Mall data set.

A. Indian Pine Site

According to the experimental design for IPS, ten percent of the samples for each class were chosen as the training set. ML-based classifiers (ML and ML_MRF) must estimate the covariance matrices of the classes before classifying all samples in the IPS hyperspectral image. These

classifiers encounter the problem of covariance matrices and poor estimations because the number of training samples in each class is less than the dimensionality. Hence, the ML and ML_MRF performance in the IPS experiment should not be compared, and is denoted as N/A.

To investigate the effects of the neighborhood systems and parameters, M and γ , Table 5 and Fig. 26 show the overall accuracies of the SCSVM and SCSVMF with the grids of $M \in \{4, 8\}$ and

$$\gamma \in \{0, 0.05, 0.1, 0.3, 0.5, 1, 10, 100, 500, 1000, 10000\}.$$

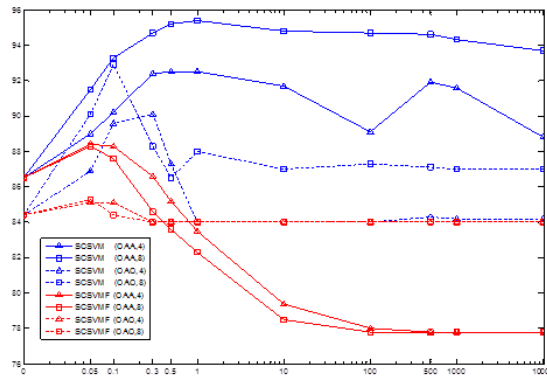


Figure 26. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the IPS data set.

Tables 6 and 7 present the validation measures of all samples in the IPS and class-specific accuracies from the best performance of k -NN classifier ($k=1$), SVM (OAO and OAA multiclass strategy), CS^4VM , SVM+EM with and without PR step, and SCSVM, which has the highest accuracy in Table 5, with and without PR step, respectively. The best overall accuracy, kappa coefficient, and average accuracy are highlighted in gray. The term “BPR” means that the performance of the classifier does not include the PR step, and “APR” means that the performance of the classifier includes the PR step.

Table 5 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in the IPS Data set.

γ	0	0.05	0.1	0.3	0.5	1	10	100	500	1000	10000
SCSVM (OAA, 4)	86.5	89.0	90.2	92.4	92.5	92.5	91.7	89.1	91.9	91.6	88.8
SCSVM (OAA, 8)	86.5	91.5	93.3	94.7	95.2	95.4	94.8	94.7	94.6	94.3	93.7
SCSVM (OAO, 4)	84.4	86.9	89.6	90.1	87.3	84.0	84.0	84.0	84.3	84.2	84.2
SCSVM (OAO, 8)	84.4	90.1	92.9	88.3	86.5	88.0	87.0	87.3	87.1	87.0	87.0
SCSVMF (OAA, 4)	86.5	88.4	88.3	86.6	85.2	83.5	79.4	78.0	77.8	77.8	77.8
SCSVMF (OAA, 8)	86.5	88.3	87.6	84.6	83.6	82.3	78.5	77.8	77.8	77.8	77.8
SCSVMF (OAO, 4)	84.4	85.1	85.1	84.0	84.0	84.0	84.0	84.0	84.0	84.0	84.0
SCSVMF (OAO, 8)	84.4	85.3	84.4	84.0	84.0	84.0	84.0	84.0	84.0	84.0	84.0

Table 6 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the IPS Data set

Classifier		Overall Accuracy (%)	Kappa Coefficient (%)	Average Accuracy (%)
ML		N/A	N/A	N/A
ML_MRF		N/A	N/A	N/A
k -NN		75.5	72.1	74.6
SVM (OAO)		84.4	82.3	85.5
SVM (OAA)		86.5	84.6	83.8
CS ⁴ VM		88.0	86.3	85.0
SVM+EM	BPR	91.3	90.0	81.6
	APR	92.8	91.8	82.5
SCSVM (OAA, $M=8$, $\gamma=1$)	BPR	95.4	94.7	94.2
	APR	95.5	94.9	94.1

Because SCSVM is a generalized version of SVM, it reverts to the original SVM when $\gamma=0$. These results show that SCSVM (OAA) can obtain a higher overall accuracy than SCSVM (OAO) in the IPS data set regardless of M and γ . Fig. 26 shows that using neighbor system ∂x_i^O generally yields better performance than using ∂x_i^F . SCSVM with $M=8$, a

second-order neighborhood system, outperforms $M=4$, a first-order neighborhood system, because the IPS is a larger spatial structure image in the original space. In hyperspectral image classification, many samples from different land-cover classes with similar spectral properties [20]-[21] affect the performance of SCSVMF. Specifically, SCSVMF performance increases only a little, and is even worse than SVM performance when γ exceeds a threshold. The highest overall SCSVM accuracy of 95.4% occurred at $M=8$ and $\gamma=1$ with the OAA multiclass strategy and the neighborhood system ∂x_i^o .

Table 7 The Class-specific Accuracies in Percentages for the IPS Data set

No.	Class Sample size	k -NN	SVM (OAO)	SVM (OAA)	CS ⁴ VM	SVM+EM		SCSVM (OAA, $M=8$, $\gamma=1$)	
						BPR	APR	BPR	APR
1	46	78.3	91.3	95.7	95.7	93.5	93.5	93.5	100.0
2	1428	64.8	78.8	86.3	88.9	86.6	89.0	82.4	86.4
3	830	62.8	82.4	79.8	81.4	89.2	90.1	93.0	94.1
4	237	54.9	95.4	77.2	79.7	97.9	100.0	97.0	100.0
5	483	89.2	90.9	91.1	91.1	93.6	94.6	95.4	95.9
6	730	94.9	93.8	94.5	93.8	97.1	98.5	95.9	97.0
7	28	85.7	96.4	85.7	85.7	0.0	0.0	100.0	100.0
8	478	96.0	84.7	97.3	97.5	97.9	98.3	86.2	88.3
9	20	40.0	60.0	45.0	45.0	5.0	0.0	95.0	100.0
10	972	74.4	89.8	85.8	85.7	87.5	90.2	97.1	98.4
11	2455	76.7	79.6	86.6	88.7	92.7	94.2	94.4	96.4
12	593	50.8	77.7	75.4	82.6	92.6	92.9	90.2	91.2
13	205	98.0	99.5	99.5	99.5	99.0	99.0	99.5	100.0
14	1265	89.7	91.8	92.6	92.6	93.3	93.8	95.9	97.3
15	386	48.7	69.4	66.8	67.6	83.7	88.3	97.2	99.2
16	93	89.2	87.1	81.7	83.9	95.7	96.8	98.9	100.0

Table 7 shows the classification maps with highest accuracies of each types of classifier for comparison. Figs. 27(a) to 27(h) show the classification maps of IPS hyperspectral image by k -NN ($k=1$), SVM (OAO), SVM (OAA), CS⁴VM, SVM+EM, and SCSVM (OAA) with $M=8$

and $\gamma = 0.1$, respectively. To conveniently compare performance, Fig. 27 (i) shows the ground truth of the IPS image.

The classification results from Table 6, Table 7, and Fig. 27 present the following findings:

1. In terms of accuracy, SCSVM (OAA) with the PR step obtained the highest overall accuracy and kappa coefficient of 95.5% and 94.9%, respectively (Table 6). However, SCSVM (OAA) without the PR step obtained the highest average accuracy of 94.2%.
2. Ten percent of each sample was selected randomly from the reference data set to serve as the training set for each class. Therefore, some classes were represented by only few training samples (i.e., there are only 3 and 2 samples for class 7 (Grass/pasture-mowed) and class 9 (Oats), respectively). This may provide an unfair representation of this class in the training process. Table 7 shows that the classification performance for class 7 with SCSVM (OAA) was better than that with k -NN, SVM (OAA), SVM (OAO), CS⁴VM, and SVM+EM. Moreover, SCSVM (OAA) achieved better performance than other classifiers in class 9. This situation was improved by SCSVM (OAA), even without the PR step, and the classification accuracies of class 7 and 9 with the PR step were 100%.
3. The classifiers with the PR step efficiently reduced some noise in the classification map, and slightly increased classification accuracy.
4. Table 6 and the classification maps in Fig. 27 based on spatial based classifiers (SVM+EM, CS⁴VM, and SCSVM) show much better results than the classifiers based on only spectral information (k -NN, SVM (OAO), and SVM (OAA)). The spatial-contextual based classifiers reduced the number of speckle-like errors, especially in areas of

Soybeans-min till, Soybeans-no till, and Corn-no till, which were the most difficult parts to classify accurately. SCSVM (OAA) achieved a great improvement in the classification map. The SCSVM (OAA) classification map (Fig. 27 (h)) was similar to the ground truth (Fig. 27 (i)) of the IPS.

5. SVM+EM also achieved sound performance on the classification maps, obtaining a 92.8% overall classification accuracy. However, this scheme relies on partitional clustering results. Hence, if the partitional clustering technique cannot accurately partition these areas, which have similar spectral properties from different classes or come from the small sample size classes (i.e., class 7 (Grass/pasture-mowed) and class 9 (Oats)), then the clustering technique will misclassify these areas into the same class (Fig. 27 (f)). If the partitional clustering technique works very well, but the standard SVM classifier cannot sensitively distinguish the pixels (e.g., different classes have similar spectral properties), then these areas will be sacrificed. Table 7 shows that SVM+EM achieved either 0% (e.g., class 7 and class 9 or low classification accuracies for small classes.
6. Since CS⁴VM is based on the OAA multiclass strategy [32], the class-specific accuracies for applying CS⁴VM are higher than or similar to those for applying SVM (OAA).

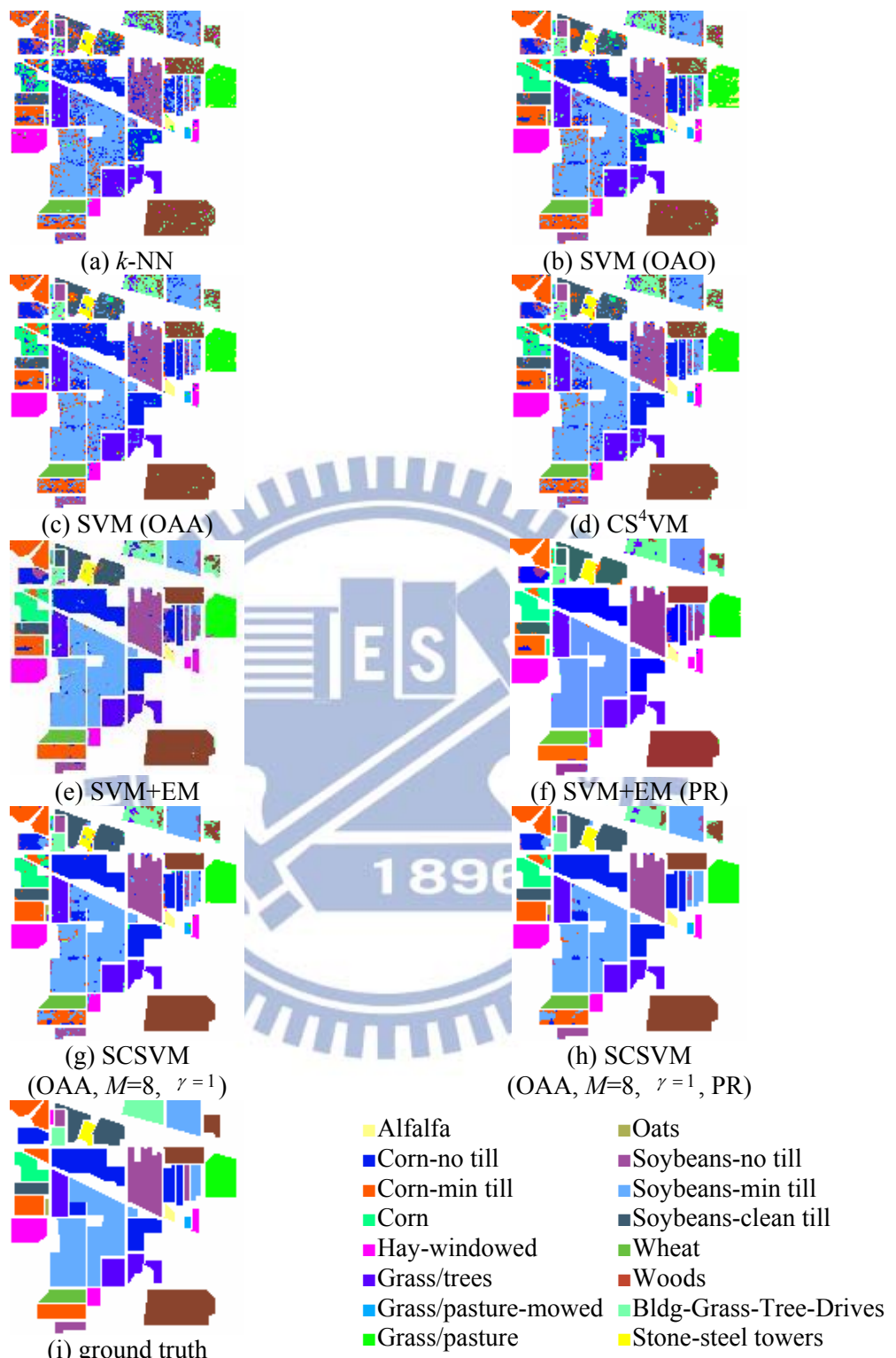


Figure 27. The classification maps of the IPS data set by the highest performance of each type classifier.

B. Washington D.C. Mall

The experiments in this study used three cases to investigate the effects of sample size on the dimensionality in the Washington D.C. Mall data set. The sample covariance matrices of ML-based classifiers, ML and ML_MRF, of case 1 and 2 will be singular. Hence, ML-based classifiers are unsuitable for case 1 and case 2, and the performance of ML-based classifiers is marked as N/A for these cases.

Similar to the IPS data set, Tables 8-10 and Fig. 28-30 respectively show the overall accuracies of SCSVM (OAA, OAO) and SCSVMF (OAA, OAO) with grids of $M \in \{4, 8\}$ and

$$\gamma \in \{0, 0.05, 0.1, 0.3, 0.5, 1, 10, 100, 500, 1000, 10000\}$$

to investigate the influence of the parameters M and γ in three cases. In case 1, the highest accuracy of 91.9% occurred at $M=4$ and $\gamma = 0.05$ with the neighborhood system in the original space (SCSVM) using the OAA multiclass strategy. However, SCSVM (OAO, 4) achieved a similar accuracy of 91.7%. The highest accuracy in case 2 is 94.1% at $M=8$ and $\gamma = 0.1$ with neighborhood system in the original space (SCSVM) and OAO multiclass strategy. However, SCSVM (OAO, 4) has a similar accuracy, at 94.0%. In case 3, the highest accuracy of 98.6% occurred at $M=4$ and $\gamma = 0.3$ with the neighborhood system in the original space (SCSVM) using the OAO multiclass strategy. The SCSVM (OAO) performance was generally better than or similar to that of SCSVM (OAA) in all three cases. Furthermore, SCSVM with $M=4$ achieved better performance than (or similar performance to) SCSVM with $M=8$. This is because the Washington D.C. Mall is an urban site without large spatial structures.

Table 8 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in Washington D.C. Mall Data set (Case 1).

γ	0	0.05	0.1	0.3	0.5	1	10	100	500	1000	10000
SCSVM (OAA, 4)	86.9	91.9	91.3	89.7	90.0	90.0	91.1	91.0	91.0	91.0	91.0
SCSVM (OAA, 8)	86.9	90.4	90.6	90.6	90.6	90.6	90.6	90.6	90.6	90.6	90.6
SCSVM (OAO, 4)	86.9	91.0	91.7	90.1	90.4	90.9	89.6	89.6	89.6	89.6	89.6
SCSVM (OAO, 8)	86.9	90.7	91.3	90.7	91.0	90.7	90.4	90.6	90.6	90.6	90.6
SCSVMF (OAA, 4)	86.9	88.3	88.3	88.1	88.0	88.1	87.6	87.7	87.7	87.7	87.7
SCSVMF (OAA, 8)	86.9	87.7	87.9	88.0	87.9	87.9	87.6	87.6	87.6	87.6	87.6
SCSVMF (OAO, 4)	86.9	86.6	86.6	86.6	86.4	85.9	86.4	86.4	86.4	86.4	86.4
SCSVMF (OAO, 8)	86.9	85.7	85.7	85.7	85.7	86.0	86.0	85.7	85.7	85.7	85.7

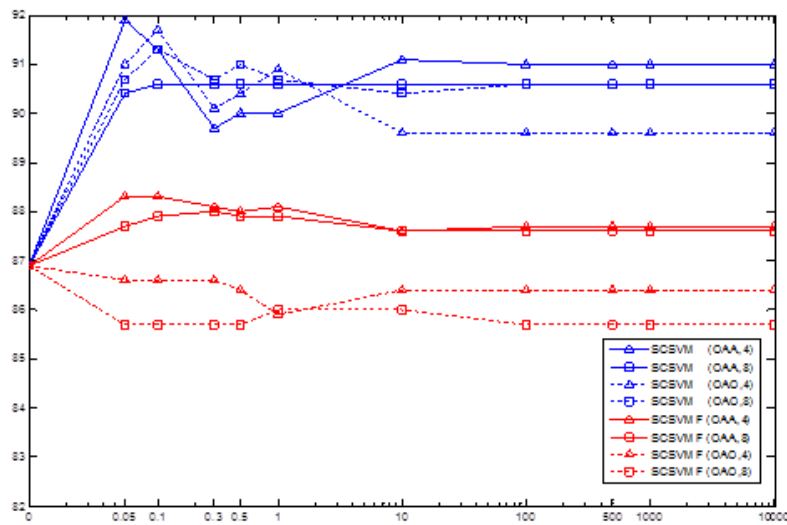


Figure 28. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the Washington D.C. Mall data set in case 1.

Figs. 29-31 show that the overall accuracy increases as the training sample size increases. Similar to the IPS data set, using many samples from

different land-cover classes, but that have similar spectral properties [20]-[21], negatively affects SCSVMF performance. That is, SCSVMF performance increases only a little, or even becomes worse than the performance of SVM, when γ exceeds a certain threshold. The performance of SCSVMF (OAO) is much worse than that of SVM (OAO).

Tables 11, 13, and 15 show the overall accuracies, kappa coefficients, and average accuracies of k -NN classifier with $k = 1$, which has the best classification performance of testing set, and SVM (OAO and OAA), CS⁴VM, SVM+EM with and without the PR step, SCSVM with and without PR step for all three cases of the Washington D.C. Mall data set. Tables 12, 14, and 16 display the class-specific accuracies of these classifiers. Because the number of testing samples is the same for every class in the Washington D.C. hyperspectral image data set (i.e., equal to 100), the overall classification accuracy and the average accuracy are identical in Tables 11, 13, and 15, and the first decimal points are all 0 in Tables 12, 14, and 16.

Similar to the IPS data set, Tables 11, 13, and 15, and Tables 12, 14, and 16, respectively compare the validation measures of SCSVM with different parameters, M and γ . These have the highest accuracies in Tables 8, 9, and 10. Fig. 31 shows a comparison of the classification maps with highest accuracies of each types of classifier in case 3.

Table 9 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in Washington D.C. Mall Data set (Case 2).

γ	0	0.05	0.1	0.3	0.5	1	10	100	500	1000	10000
SCSVM (OAA, 4)	89.0	93.4	93.3	93.4	93.7	93.7	93.4	93.3	93.1	93.1	93.1
SCSVM (OAA, 8)	89.0	91.9	92.7	93.1	92.9	93.0	92.9	92.4	92.4	92.4	92.4
SCSVM (OAO, 4)	88.6	92.4	93.6	94.0	93.7	92.4	91.0	90.4	90.4	90.4	90.7
SCSVM (OAO, 8)	88.6	93.0	94.1	92.4	92.6	92.4	92.0	91.9	91.9	91.9	91.9
SCSVMF (OAA, 4)	89.0	88.9	88.9	89.1	89.3	89.3	89.1	89.3	89.3	89.3	89.3
SCSVMF (OAA, 8)	89.0	88.9	88.7	88.7	88.9	89.0	88.9	88.9	88.9	88.9	88.9
SCSVMF (OAO, 4)	88.6	88.6	88.6	88.9	88.9	88.3	88.0	88.1	88.0	88.1	88.1
SCSVMF (OAO, 8)	88.6	88.1	88.3	88.3	88.1	87.9	88.0	88.0	88.0	88.0	88.0

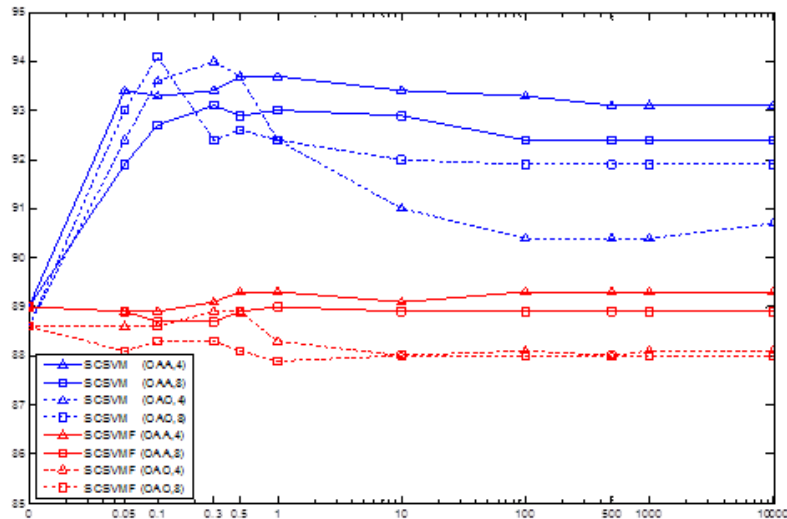
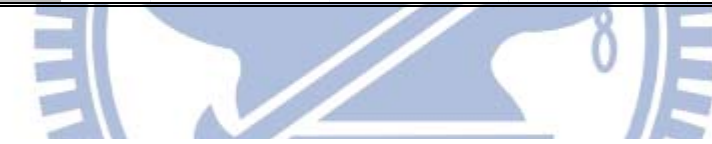


Figure 29. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the Washington D.C. Mall data set in case 2.

Table 10 The Overall Accuracies in Percentages of SCSVM (OAA, M), SCSVM (OAO, M), SCSVMF (OAA, M), and SCSVMF (OAO, M) with Different Parameters M , the Size of the Neighborhood System, and γ in Washington D.C. Mall Data set (Case 3).

γ	0	0.05	0.1	0.3	0.5	1	10	100	500	1000	10000
SCSVM (OAA, 4)	93.7	95.3	95.6	95.9	96.1	97.4	96.9	95.7	95.3	95.3	95.3
SCSVM (OAA, 8)	93.7	94.4	94.6	95.1	95.3	95.3	95.4	95.5	94.4	94.4	94.4
SCSVM (OAO, 4)	94.3	94.9	97.3	98.6	98.1	97.3	96.1	95.4	96.1	96.1	96.1
SCSVM (OAO, 8)	94.3	95.9	97.3	96.9	96.0	96.0	95.0	95.4	95.1	95.1	95.1
SCSVMF (OAA, 4)	93.7	94.3	94.4	94.3	94.1	94.3	94.9	94.1	94.1	94.1	94.1
SCSVMF (OAA, 8)	93.7	94.1	94.4	94.4	94.4	94.6	94.4	94.4	94.4	94.4	94.4
SCSVMF (OAO, 4)	94.3	93.7	94.0	93.9	93.6	93.4	93.4	93.1	93.1	93.1	93.1
SCSVMF (OAO, 8)	94.3	93.9	94.0	93.9	93.7	93.4	93.6	93.4	93.4	93.4	93.4

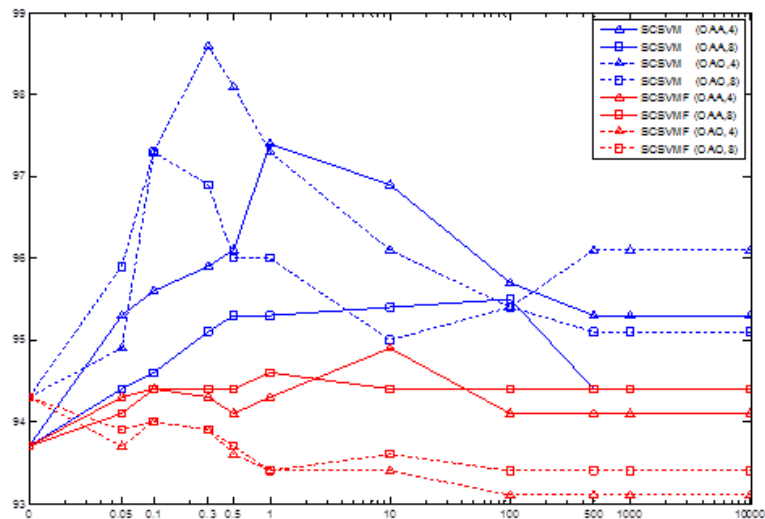


Figure 30. The overall accuracies in percentages of the experimental classifiers, SCSVM and SCSVMF, for the Washington D.C. Mall data set in case 3.

Table 11 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the Washington D.C. Mall Data set (Case1).

Classifier		Overall Accuracy	Kappa Coefficient	Average Accuracy
ML		N/A	N/A	N/A
ML_MRF		N/A	N/A	N/A
k -NN		85.6	83.2	85.6
SVM (OAO)		86.9	84.7	86.9
SVM (OAA)		86.9	84.7	86.9
CS4VM		87.7	85.7	87.7
SVM+EM	BPR	82.0	79.0	82.0
	APR	80.3	77.0	80.3
SCSVM (OAA, $M=4$, $\gamma = 0.05$)	BPR	91.9	90.5	91.9
	APR	92.0	90.6	92.0

Table 12 The Class-Specific Accuracies in Percentages for the Washington D.C. Mall Data set in Case 1

Class No.	Sample size	k -NN	SVM (OAO)	SVM (OAA)	CS ⁴ VM	SVM+EM		SCSVM(OAA, $M=4$, $\gamma = 0.05$)	
						BPR	APR	BPR	APR
1	100	79.0	78.0	81.0	86.0	80.0	83.0	92.0	92.0
2	100	82.0	94.0	93.0	99.0	93.0	92.0	100.0	100.0
3	100	58.0	66.0	60.0	57.0	62.0	60.0	93.0	91.0
4	100	98.0	93.0	94.0	96.0	100.0	100.0	98.0	97.0
5	100	94.0	95.0	95.0	98.0	54.0	51.0	83.0	86.0
6	100	90.0	90.0	91.0	85.0	91.0	83.0	85.0	84.0
7	100	98.0	92.0	94.0	93.0	94.0	93.0	92.0	94.0

Table 13 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the Washington D.C. Mall Data set (Case 2).

Classifier		Overall Accuracy	Kappa Coefficient	Average Accuracy
ML		N/A	N/A	N/A
ML_MRF		N/A	N/A	N/A
k -NN		87.1	85.0	85.6
SVM (OAO)		88.6	86.7	88.6
SVM (OAA)		89.0	87.2	89.0
CS4VM		89.4	87.7	89.4
SVM+EM	BPR	86.7	84.5	86.7
	APR	84.6	82.0	84.6
SCSVM (OAO, $M=8$, $\gamma = 0.1$)	BPR	94.1	93.2	94.1
	APR	94.1	93.2	94.1

Table 14 The Class-Specific Accuracies in Percentages for the Washington D.C. Mall Data set in Case 2.

Class No.	Sample size	k -NN	SVM (OAO)	SVM (OAA)	CS ⁴ VM	SVM+EM		SCSVM(OAO, $M=8, \gamma = 0.1$)	
						BPR	APR	BPR	APR
1	100	86.0	95.0	95.0	97.0	98.0	99.0	98.0	98.0
2	100	84.0	95.0	95.0	100.0	96.0	95.0	100.0	100.0
3	100	66.0	60.0	59.0	54.0	58.0	56.0	82.0	82.0
4	100	98.0	98.0	98.0	98.0	100.0	100.0	100.0	100.0
5	100	96.0	90.0	91.0	98.0	70.0	66.0	98.0	98.0
6	100	87.0	91.0	92.0	85.0	91.0	83.0	84.0	84.0
7	100	93.0	91.0	93.0	94.0	94.0	93.0	97.0	97.0

Table 15 The Overall Accuracies, Kappa Coefficients, and Average Accuracies in Percentages of the Experimental Classifiers for the Washington D.C. Mall Data set (Case 3).

Classifier		Overall Accuracy	Kappa Coefficient	Average Accuracy
ML		94.1	93.2	94.1
ML_MRF		96.7	96.2	96.7
k -NN		94.4	93.5	94.4
SVM (OAO)		94.3	93.3	94.3
SVM (OAA)		93.7	92.7	93.7
CS4VM		94.1	93.2	94.1
SVM+EM	BPR	94.6	93.7	94.6
	APR	92.9	91.7	92.9
SCSVM (OAO, $M=4, \gamma = 0.3$)	BPR	98.6	98.3	98.6
	APR	98.4	98.1	98.4

Table 16 The Class-Specific Accuracies in Percentages for the Washington D.C. Mall Data set in Case 3

No.	Class Sample size	ML	ML_MRF	<i>k</i> -NN	SVM(OAO)	SVM(OAA)	CS ⁴ VM	SVM+EM		SCSVM (OAO, $M=4$, $\gamma = 0.3$)	
								BPR	APR	BPR	APR
1	100	99.0	100.0	95.0	97.0	96.0	99.0	98.0	99.0	99.0	99.0
2	100	99.0	99.0	96.0	99.0	99.0	99.0	93.0	92.0	100.0	100.0
3	100	90.0	94.0	80.0	78.0	75.0	76.0	85.0	85.0	99.0	100.0
4	100	98.0	99.0	100.0	98.0	98.0	98.0	100.0	100.0	99.0	100.0
5	100	93.0	99.0	99.0	100.0	100.0	100.0	98.0	97.0	99.0	100.0
6	100	85.0	90.0	93.0	90.0	90.0	89.0	91.0	83.0	97.0	94.0
7	100	95.0	96.0	98.0	98.0	98.0	98.0	97.0	94.0	97.0	96.0

The classification results in Tables 11-16 and Fig. 31 reveal the following findings:

1. SCSVM obtained the highest classification accuracies of the testing set in terms of overall accuracy, kappa coefficient, and average classification for all cases. In case 1, SCSVM (OAA) with $M=4$, $\gamma=0.05$ and the PR step achieved the best classification accuracy. The overall accuracy, kappa coefficient, and average classification were 92.0%, 90.6%, and 92.0%, respectively. In case 2, SCSVM (OAA) with $M=8$, $\gamma=0.1$ achieved the best classification accuracy. The overall accuracy, kappa coefficient, and average classification were 94.1%, 93.2%, and 94.1%, respectively. In case 3, SCSVM (OAO) with $M=4$, $\gamma=0.3$ and without the PR step achieved the best classification accuracy. The overall accuracy, kappa coefficient, and average classification were 98.6%, 98.3%, and 98.6%, respectively. The accuracies increased in all classifiers as the training sample size increased.
2. Because the Washington D.C. image is an urban site image, some areas in the image are small spatial structures and some areas are large spatial structures. The PR step did not work well for small spatial structure areas (e.g. class 6 (trail) in Fig. 31). Tables 12, 14, and 16 show that the accuracy decreased upon applying the PR step in class 6 (trail). However, the PR step improved some noisy pixels in large structure areas (e.g., class 4 (water)), with the exception of case 1, which encountered the small sample size problem.
3. The image contains too many types of roofs. Hence, when the training sample size is small (e.g., in class 3 (roof)) some types of roofs may not be chosen as training samples. For this reason, the classification accuracies and maps of this class are poor in case 1 and case 2 (Tables

12 and 14) regardless of the classifier. In the general case (case 3), these classes are identified more accurately (Table 16).

4. Fig. 32 shows classified images of SCSVM (OAO) and SCSVM (OAA) with $M=4$ and $\gamma = 0, 0.1, \text{ and } 0.3$. These images reveal the effects of parameter γ on the classified image. As the gamma increased, the classified image exhibited more homogeneous groups of pixels.
5. Most of the spatial based classifiers (ML_MRF, CS⁴VM, and SCSVM) achieved better classification performance than the spectral-information-only-based classifiers (ML, k -NN, SVM). The exception here is SVM+EM in cases 1 and 2, because SVM+EM is particularly suitable for classifying images with large spatial structures [31]. The drawback of SVM+EM is that when including spatial information from the segmentation map or from the closest neighborhoods in a classifier, small spatial structures may be assimilated with larger neighboring structures if the spectral responses are not significantly different [31]. Hence, SVM+EM is not really suitable for the small areas of the Washington D.C. image. However, the classification accuracy and maps of class 4 (water), which is a large structure, can be improved, and the class-specific accuracy of the class (water) is 100% (Tables 12, 14, and 16) for all three cases.
6. The CS⁴VM classifier is based on the OAA multiclass strategy. This classifier achieved slightly better classification accuracy than SVM (OAA) (see Tables 11, 13, and 15), but SCSVM still achieved better performance than CS⁴VM. However, the CS⁴VM distinguished some areas better than SCSVM.

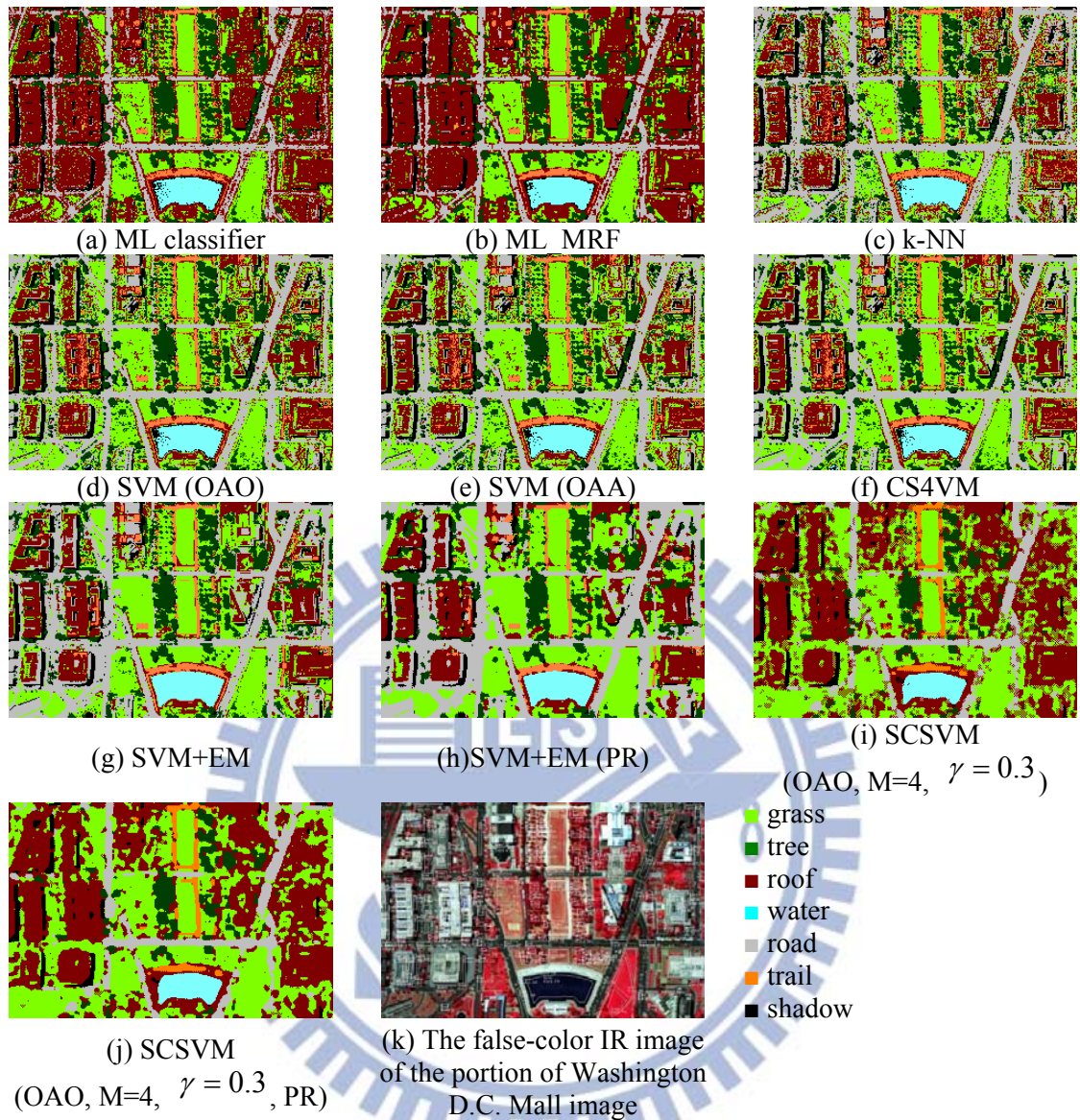


Figure 31. The classification maps of a portion of the Washington D.C. data set (case 3) by the highest performance of each type classifier.

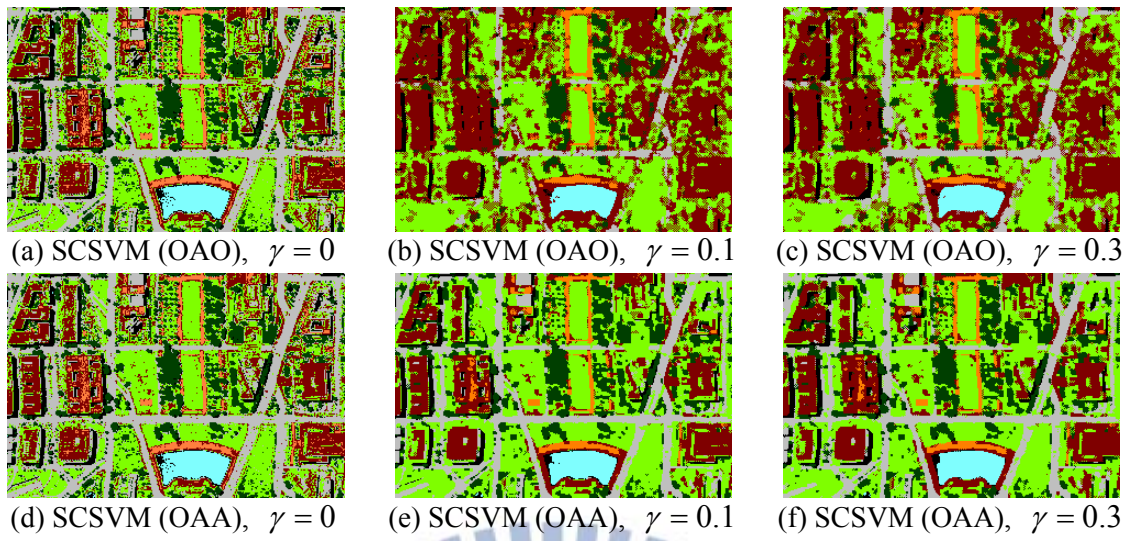
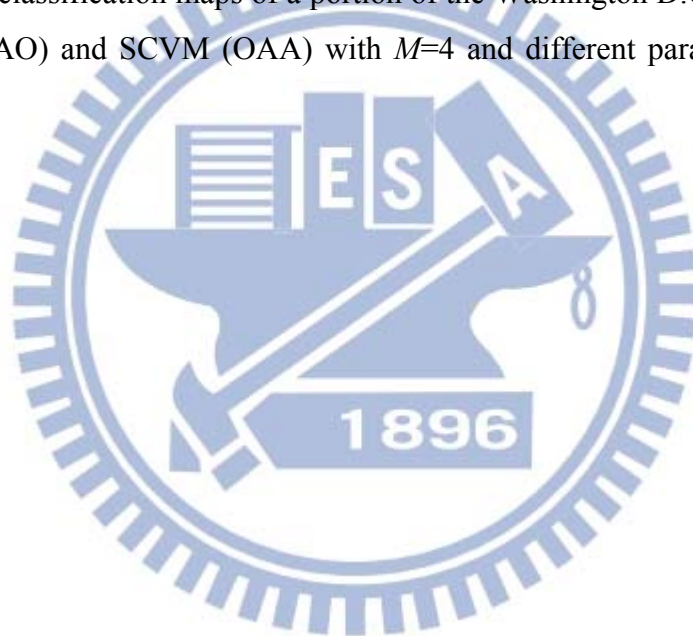


Figure 32. The classification maps of a portion of the Washington D.C. data set (case 3) of SCSVM (OAO) and SCVM (OAA) with $M=4$ and different parameters $\gamma=0, 0.1,$ and 0.3 .



6. Conclusion

This study proposes a clustering algorithm, called FLDC, and two kinds of spatial-contextual support vector machines (SCSVMs). FLDC is based on the Fisher criterion composed of the fuzzy between- and within-cluster scatter matrices extended from LDA. Experimental results with both synthetic and real data indicate that the proposed clustering algorithm outperformed the KMS, KMD, FCM, GK, GG, PCM, FPCM, PFCM, FCS, FSMM, and FMSFA algorithms.

The results of clustering synthetic data sets reveal that FLDC only worked well when the distribution of clusters showed a normal distribution. Hence, future research should extend FLDC using kernel tricks, that is, a clustering algorithm based on an unsupervised version of kernel-based LDA for non-normal data sets.

Another direction for future research is to show that the proposed optimization problem is non-convex and nonlinear. Although the proposed methods work well, the optimal solution may fall into a local minimum, and the interior-point optimization method is time consuming. Thus, it is necessary to find a more efficient algorithm for solving such problems.

The number of clusters is an important factor in all clustering algorithms. Future research should develop or choose an appropriate criterion for FLDC, [Akaike and Bayesian information criteria (AIC and BIC)], to determine the number of clusters.

For SCSVMs, results show that a SCSVM based on the neighborhood system in the original space can overcome similar spectral properties. SCSVM modifies the decision function and the constraints of SVM based on spatial-contextual information. A PR step consisting of a fixed-window-based postfiltering was employed to reduce the remaining

noise in the classification map. The experiments in this study compared and analyzed the effects of different types of classifiers on the classification accuracy and classification map of the proposed SCSVM, ML classifier, ML-MRF classifier, k -NN classifier, a standard supervised SVM, a CS⁴VM, and SVM+EM.

The experimental results obtained from two different hyperspectral image data sets, the Indian Pine site (a mixed forest/agricultural site in Indiana) and the Washington D.C. Mall hyperspectral image (an urban site in Washington D.C.), confirm that the proposed SCSVM improves the classification accuracies and kappa coefficients.

This discussion leads to the following conclusions about SCSVMs.

1. SCSVM (OAA) performs better than or similar to SCSVM (OAO) in the IPS data set. The classification map of IPS data set obtained from SCSVM (OAA) with the PR step (Fig. 27 (h)) is very close to the ground truth, and the SCSVM classification accuracy and kappa coefficient are 95.5% and 94.9%, respectively. However, in the Washington D.C. Mall data set, SCSVM (OAO) performs better than or similar to SCSVM (OAA), and SCSVMF (OAA) performs better than or similar to SCSVMF (OAO).
2. This study shows that selecting a suitable spatial parameter γ improves SCSVM performance, and the best choice of γ becomes larger as the training sample size increases. That is, γ has a significant influence on performance, especially for the SCSVM (OAA).
3. The computational cost of the learning phase in the proposed SCSVM is slightly higher than that of the standard SVM in each round. From a theoretical viewpoint, a standard supervised SVM is a special case of

SCSVM if the parameter γ is equal to 0. However, CS⁴VM requires a huge semi-sample set from the neighborhoods of each training sample in the objective function. Hence, the computational cost of the CS⁴VM learning phase is slightly higher than that of SCSVM learning phase. This is because SCSVM only uses the same training sample in the objective function in each round. For example, in the IPS data set experiment, the training phase of a supervised SVM (OAA) took about 7.566s on a PC with an Intel Core 2 Duo CPU at 2.4 GHz and a 4-Gb DDR2 RAM. The training phase of SCSVM (OAA) took about 7.909s on the same machine, but the training phase of CS⁴VM required about 185.56s.

4. The SVM+EM method is particularly suitable for classifying images with large spatial structures (e.g., the IPS image) when the spectral responses of different classes are dissimilar and the classes contain a comparable number of pixels. However, most real data does not always satisfy this condition (e.g., the Washington D.C. Mall image). Hence, SVM+EM is not suitable for all situations. In the SCSVM classifier, the spatial neighborhood system can be modified according to the spatial structures of different data sets.

References

- [1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, New York: Springer-Verlag, 2001.
- [2] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd edition, Academic Press, 2006.
- [3] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. London, U.K.: Wiley, Nov. 2009.
- [4] C.-H. Li, B.-C. Kuo, and C.-T. Lin, "LDA-based clustering algorithm and its application to an unsupervised feature extraction," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 1, pp.152-163, Feb. 2011.
- [5] B. Balasko, J. Abonyi, and B. Feil, *Fuzzy clustering and data analysis toolbox for use with Matlab*, Available from: <<http://www.fmt.vein.hu/softcomp>>.
- [6] C.-T. Lin and C.-S. George Lee, *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems*. Prentice Hall, 1996.
- [7] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
- [8] P. J. Rouseeuw, L. Kaufman, and E. Trauwaert, "Fuzzy clustering using scatter matrices," *Computational Statistics & Data Analysis*, vol. 23, pp. 135-151, 1996.
- [9] D.E. Gustafson and W.C. Kessel, "Fuzzy clustering with fuzzy covariance matrix," In *Proceedings of the IEEE CDC*, San Diego, pp. 761-766, 1979.
- [10] N.R. Pal, K. Pal, and J.C. Bezdek, "A mixed c-means clustering model," *IEEE International Conference on Fuzzy Systems*, pp. 11-21, 1997
- [11] N.R. Pal, K. Pal, J.M. Keller, J.C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517-530, 2005.
- [12] J.C. Bezdek and J.C. Dunn, "Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions," *IEEE Transactions on Computers*, pp. 835-838, 1975.
- [13] S. Chatzis and T. Varvarigou, "Robust fuzzy clustering using mixtures of student's-t distributions," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1901-1905, October 2008.
- [14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA: Academic, 1990.
- [15] K.-L. Wu, J. Yu, and M.-S. Yang, "A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests," *Pattern Recognition Letters*, vol. 26, pp. 639-652, 2005.
- [16] Z. Yin, Y. Tang, F. Sun, and Z. Sun, "Fuzzy clustering with novel separable criterion," *Tsinghua Science & Technology*, vol. 11, no. 1, pp. 50-53, Feb. 2006.
- [17] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

- [18] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp.4085-4098, Nov. 2010.
- [19] G. Camps-Valls, T. V. B. Maratheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp.3044-3054, Oct. 2007.
- [20] Q. Jackson and D. A. Landgrebe, "Adaptive Bayesian contextual classification based on Markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 11, pp. 2454-2463, 2002.
- [21] B.-C. Kuo, C.-H. Chuang, C.-S. Huang, and C.-C. Hung, "A nonparametric contextual classification based on Markov random fields," *WHISPERS '09 - 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009.
- [22] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, John Wiley and Sons, Hoboken, NJ: Chichester, 2003.
- [23] B.E. Boser, I.M. Guyon, and V.N. Vapnik. "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp.144-152, 1992.
- [24] V.N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2001.
- [25] Y.P. Zhao and J.G. Sun, "A fast method to approximately train hard support vector regression," *Neural Networks*, vol. 23, no. 10, pp. 1276-1285, Dec. 2010.
- [26] K. Ersahin, I. G. Cumming, and R. K. Ward, "Segmentation and classification of polarimetric SAR data using spectral graph partitioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 1, pp.164-174, Jan. 2010.
- [27] G. Camps-Valls, N. Shervashidze, and K. M. Borgwardt, "Spatio-spectral remote sensing image classification with graph kernels," *IEEE Transactions on Geoscience and Remote Sensing Letter*, vol. 7, no. 4, pp.741-745, Oct. 2010.
- [28] M. Fauvel, "Spectral and spatial methods for the classification of urban remote sensing data," *Ph.D. dissertation*, Grenoble Inst. Technol., Grenoble, France, 2007.
- [29] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, pp. 309-320, Feb. 2001.
- [30] M. Fauvel, J. Chanussot, J. A. Benediktsson, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 3804-3814, Oct. 2008.
- [31] Y. Tarabalka, J.A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitionial clustering techniques." *IEEE Transactions on Geoscience and Remote Sensing*, vol.47, no.8, pp. 2973-2987, Aug. 2009.

- [32] L. Bruzzone and C. Persello, "A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, issue 7, pp. 2142-2154, 2009.
- [33] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, issue 1, pp. 93- 97, 2006.
- [34] S. T. John, and C. Nello, *Kernel Methods for Pattern Analysis*, Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [35] F. Melgani, and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778-1790, Aug. 2004.
- [36] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [37] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351-1362, Jun. 2005.
- [38] M. Fauvel, J. Chanussot, and J.A. Benediktsson, "Evaluation of kernels for multiclass classification of hyperspectral remote sensing data," in *Proceedings of ICASSP*, pp. II-813–II-816, May 2006.
- [39] C.-H. Li, B.-C. Kuo, C.-T. Lin, and C.-S. Huang, "A spatial-contextual support vector machine for remotely sensed image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1-16, August 2011.
- [40] A.K. Jain, R.P.W. Duin, and J.C. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.
- [41] E. Alpaydin, *Introduction to Machine Learning*. MIT Press 2004.
- [42] P.F. Hsieh, D.S. Wang, and C.W. Hsu "A linear feature extraction for multi-class classification problems based on class mean and covariance discriminant information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 223-235, Feb. 2006.
- [43] S.J. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in kernel Fisher discriminant analysis," in *International Conference on Machine Learning*, Pittsburgh, PA, 2006, pp. 465-472.
- [44] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Transactions on Neural Networks*, vol. 16, no. 2, pp. 460-474, Mar. 2005.
- [45] P.S. Szczepaniak, P.J.G. Lisboa, and J. Kacprzyk, *Fuzzy Systems in Medicine*. Physica-Verlag Heidelberg New York, 2000.
- [46] N. Pal, K. Pal, J. Keller, J. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, 13 (4) (2005) 517–530.
- [47] B.-C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, pp.1096-1105, May 2004.

- [48] B.-C. Kuo, D. A. Landgrebe, L.-W. Ko, and C.-H. Pai, "Regularized feature extractions for hyperspectral data classification," *International Geoscience and Remote Sensing Symposium (IGARSS)*, Toulouse, France, July 21–25, 2003.
- [49] R.A. Waltz, J.L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Mathematical Programming*, vol 107, no. 3, pp. 391-408, 2006.
- [50] D.G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 3rd ed., New York, NY: Springer, 2009.
- [51] S. Chatzis, http://web.mac.com/soteri0s/Sotirios_Chatzis/Software.html
- [52] R. Krishnapuram and J. Keller, "A possibilistic Approach to Clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, May 1993.
- [53] L.I. Kuncheva and D.P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1798-1808, 2006.
- [54] L.I. Kuncheva, http://www.bangor.ac.uk/~mas00a/activities/artificial_data.htm
- [55] C.L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases," 1998, <http://www.ics.uci.edu/~mllearnMLRepository.html>
- [56] S. Chatzis and T. Varvarigou, "Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using t-Distributions," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 3, pp. 505-517, June 2009.
- [57] G. McLachlan, R. Bean, and L. B.-T. Jones, "Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution," *Computational Statistics & Data Analysis*, vol. 51, no. 11, pp. 5327-5338, 2007.
- [58] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 508-516, Aug. 2005.
- [59] Z. Ghahramani and G. Hinton, "The EM algorithm for mixtures of factor analyzers," Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep. CRGTR-96-1, 1997.
- [60] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer Learning Revisited: a Stepwise Procedure for Building and Training a Neural Network," in *J. Fogelman, editor, Neurocomputing: Algorithms, Architectures and Applications. Springer-Verlag*, 1990.
- [61] C.-W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transaction on Neural Networks*, vol. 13, issue 2, pp. 415-425, 2002.
- [62] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. "Comparison of classifier methods: a case study in handwriting digit recognition." in *Proceedings International Conference on Pattern Recognition*, pp. 77-87, 1994.
- [63] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel Nonparametric Weighted Feature Extraction for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp.1139-1155, April 2009.

- [64] J.-M. Yang, P.-T. Yu, and B.-C. Kuo, "A Nonparametric Feature Extraction and Its Application to Nearest Neighbor Classification for Hyperspectral Image Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp.1279-1293, March 2010.
- [65] J.-M. Yang, B.-C. Kuo, P.-T. Yu, and C.H. Chuang, "A Dynamic Subspace Method for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp.2840-2853, July 2010.
- [66] A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition*, 3rd Edition, John Wiley & Sons, Ltd, Chichester, UK, 2011.

