

Member, IEEE, and J. A. Kittler, Senior Member, IEEE

Abstract—This paper addresses the problem of the classification of hyperspectral remote sensing images by support vector machines (SVMs). First, we propose a theoretical discussion and experimental analysis aimed at understanding and assessing the potentialities of SVM classifiers in hyperdimensional feature spaces. Then, we assess the effectiveness of SVMs with respect to conventional feature-reduction-based approaches and their performances in hypersubspaces of various dimensionalities. To sustain such an analysis, the performances of SVMs are compared with those of two other nonparametric classifiers (i.e., radial basis function neural networks and the K-nearest neighbor classifier). Finally, we study the potentially critical issue of applying binary SVMs to multiclass problems in hyperspectral data. In particular, four different multiclass strategies are analyzed and compared: the one-against-all, the one-against-one, and two hierarchical tree-based strategies. Different performance indicators have been used to support our experimental studies in a detailed and accurate way, i.e., the classification accuracy, the computational time, the stability to parameter setting, and the complexity of the multiclass architecture. The results obtained on a real Airborne Visible/Infrared Imaging Spectroradiometer hyperspectral dataset allow to conclude that, whatever the multiclass strategy adopted, SVMs are a valid and effective alternative to conventional pattern recognition approaches (feature-reduction procedures combined with a classification method) for the classification of hyperspectral remote sensing data.

Index Terms—Classification, feature reduction, Hughes phenomenon, hyperspectral images, multiclass problems, remote sensing, support vector machines (SVMs).

Remote sensing of the Earth's surface is a very important activity for many applications, such as land use and land cover classification, environmental monitoring, and resource management. In the last few years, the availability of hyperspectral remote sensing data has increased significantly, leading to a growing interest in the development of new classification methods. Support vector machines (SVMs) have emerged as a powerful tool for this task, due to their ability to handle high-dimensional data and their robustness to overfitting. In this paper, we investigate the performance of SVMs in the classification of hyperspectral remote sensing data, comparing them with conventional feature-reduction-based approaches and other nonparametric classifiers. We also study the issue of applying binary SVMs to multiclass problems, analyzing four different strategies: one-against-all, one-against-one, and two hierarchical tree-based strategies. The results show that SVMs are a valid and effective alternative to conventional pattern recognition approaches for the classification of hyperspectral remote sensing data.

The Hughes phenomenon is a well-known problem in the classification of hyperspectral data, where the performance of a classifier decreases as the number of features increases. This is due to the fact that the number of samples available for training and testing is limited, and the variance of the estimated parameters increases as the dimensionality of the feature space increases. SVMs have been shown to be robust to this phenomenon, as they do not require the estimation of parameters for each feature. In this paper, we compare the performance of SVMs with that of other classifiers in the presence of the Hughes phenomenon. The results show that SVMs are able to maintain a high classification accuracy even in high-dimensional feature spaces, while other classifiers suffer from a significant performance drop.

The Hughes phenomenon is a well-known problem in the classification of hyperspectral data, where the performance of a classifier decreases as the number of features increases. This is due to the fact that the number of samples available for training and testing is limited, and the variance of the estimated parameters increases as the dimensionality of the feature space increases. SVMs have been shown to be robust to this phenomenon, as they do not require the estimation of parameters for each feature. In this paper, we compare the performance of SVMs with that of other classifiers in the presence of the Hughes phenomenon. The results show that SVMs are able to maintain a high classification accuracy even in high-dimensional feature spaces, while other classifiers suffer from a significant performance drop.

The Hughes phenomenon is a well-known problem in the classification of hyperspectral data, where the performance of a classifier decreases as the number of features increases. This is due to the fact that the number of samples available for training and testing is limited, and the variance of the estimated parameters increases as the dimensionality of the feature space increases. SVMs have been shown to be robust to this phenomenon, as they do not require the estimation of parameters for each feature. In this paper, we compare the performance of SVMs with that of other classifiers in the presence of the Hughes phenomenon. The results show that SVMs are able to maintain a high classification accuracy even in high-dimensional feature spaces, while other classifiers suffer from a significant performance drop.

[Musical notation: Treble clef, 2/4 time signature, key signature of one flat. The staff contains a series of notes and rests, with some notes beamed together. The piece concludes with a double bar line and repeat dots.]

et al.

[Musical notation: Treble clef, 2/4 time signature, key signature of one flat. The staff contains a series of notes and rests, with some notes beamed together. The piece concludes with a double bar line and repeat dots.]

one

[Musical notation: Treble clef, 2/4 time signature, key signature of one flat. The staff contains a series of notes and rests, with some notes beamed together. The piece concludes with a double bar line and repeat dots.]

A. SVM Mathematical Formulation

1) Linear SVM: Linearly Separable Case:

Let $\mathbf{x}_i \in \mathbb{R}^d$ ($i = 1, 2, \dots, N$), $y_i \in \{-1, +1\}$ be the training data. We seek a hyperplane defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$ such that the data points are separated by the hyperplane.

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b. \quad (1)$$

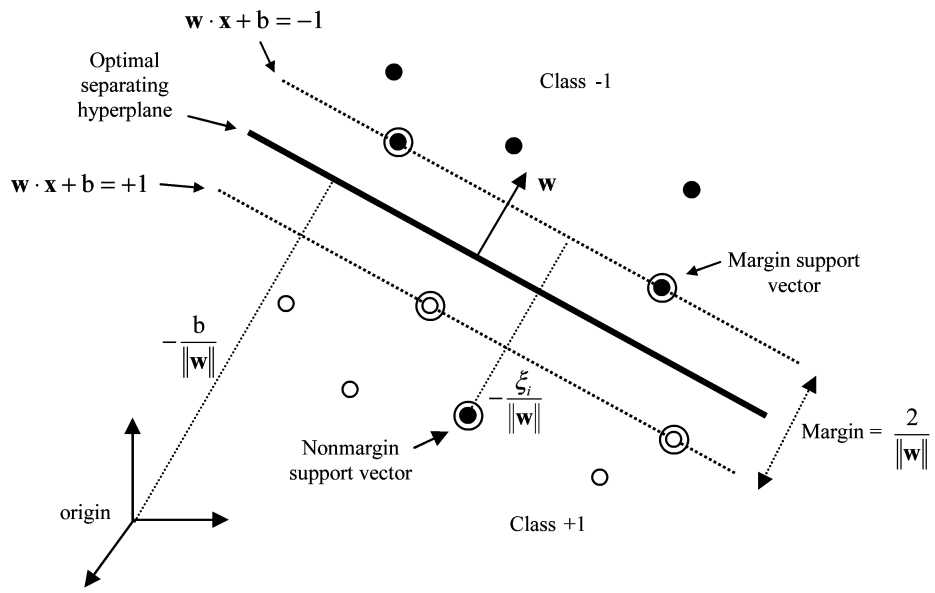
$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 0, \quad i = 1, 2, \dots, N. \quad (2)$$

$$\min_{i=1,2,\dots,N} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1. \quad (3)$$

$$\frac{1}{\|\mathbf{w}\|} (\mathbf{w} \cdot \mathbf{x}_i + b) \geq \frac{1}{\|\mathbf{w}\|}. \quad (4)$$

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 \\ y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \dots, N \end{cases} \quad (5)$$

$$\begin{cases} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ \sum_{i=1}^N \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, N. \end{cases} \quad (6)$$



α_i ($i = 1, 2, \dots, N$)
 support vectors

$$f(x) = \sum_{i \in S} \alpha_i y_i (x_i \cdot x) + b$$

2) Linear SVM: Linearly Nonseparable Case:
 $1/\|w\|$

$$\Psi(w, \xi) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^N \xi_i$$

ξ_i slack variables

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

$$\xi_i \geq 0, \quad i = 1, 2, \dots, N.$$

3) Nonlinear SVM: Kernel Method:

$\Phi(\cdot)$ $\Phi(x) \in \mathbb{R}^{d'} \ (d' > d)$
 $w \in \mathbb{R}^{d'} \quad b \in \mathbb{R}$
 $(x_i \cdot x_j)$
 $[\Phi(x_i) \cdot \Phi(x_j)]$
 $\Phi(x)$
 $K(x_i, x) = \Phi(x_i) \cdot \Phi(x).$

$$\begin{cases} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \sum_{i=1}^N \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N. \end{cases}$$

The decision function is given by

$$f(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

For the linear kernel, the decision function is given by

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2)$$

For the polynomial kernel, the decision function is given by

$$K(\mathbf{x}_i, \mathbf{x}) = [\mathbf{x}_i \cdot \mathbf{x} + 1]^p.$$

Support Vector Machines (SVMs) are a class of supervised learning models that are designed to find the optimal decision boundary between classes in a high-dimensional feature space. The decision function is given by $f(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$, where \mathbf{x}_i are the support vectors, α_i are the Lagrange multipliers, y_i are the class labels, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function, and b is the bias term. The kernel function maps the input data into a higher-dimensional space where the classes are linearly separable. The decision function is then used to classify new data points based on their position relative to the decision boundary.

B. SVMs in Hyperspectral Feature Spaces

Hyperspectral data is characterized by a large number of narrow spectral bands, which can be used to distinguish between different materials and objects. SVMs are well-suited for this task because they can handle high-dimensional data and find the optimal decision boundary between classes. The decision function is given by $f(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$, where \mathbf{x}_i are the support vectors, α_i are the Lagrange multipliers, y_i are the class labels, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function, and b is the bias term. The kernel function maps the input data into a higher-dimensional space where the classes are linearly separable. The decision function is then used to classify new data points based on their position relative to the decision boundary.

The decision function is given by $f(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$, where \mathbf{x}_i are the support vectors, α_i are the Lagrange multipliers, y_i are the class labels, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function, and b is the bias term. The kernel function maps the input data into a higher-dimensional space where the classes are linearly separable. The decision function is then used to classify new data points based on their position relative to the decision boundary.

$$R_V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)}$$

The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.

The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.

The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.

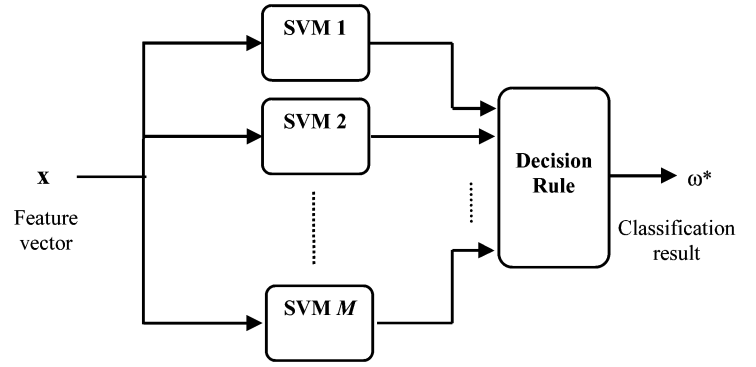
The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.

The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.

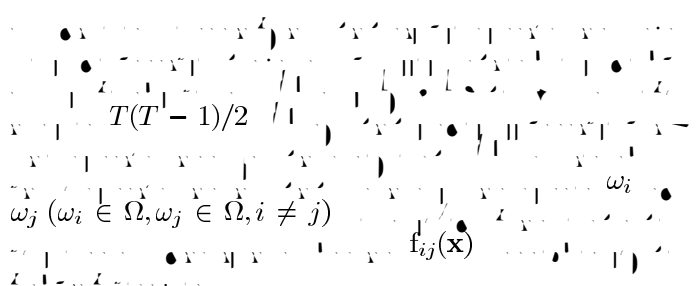
The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.

If you possess a limited amount of information to solve a problem, try solving it directly and never solve a more general problem as an intermediate step. The available information may be sufficient for a direct solution, though insufficient to solve a more general intermediate problem.

The volume of a d -dimensional sphere is given by $V = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} R^d$, where R is the radius of the sphere.



Let $\mathbf{X} = \{x_1, \dots, x_T\}$ be a set of T feature vectors, M SVMs, $T(T-1)/2$



$$\begin{cases} \Omega_A = \omega_i \\ \Omega_B = \omega_j. \end{cases} \quad (1)$$

$\omega_i \in \Omega$ $S_i(\mathbf{x})$

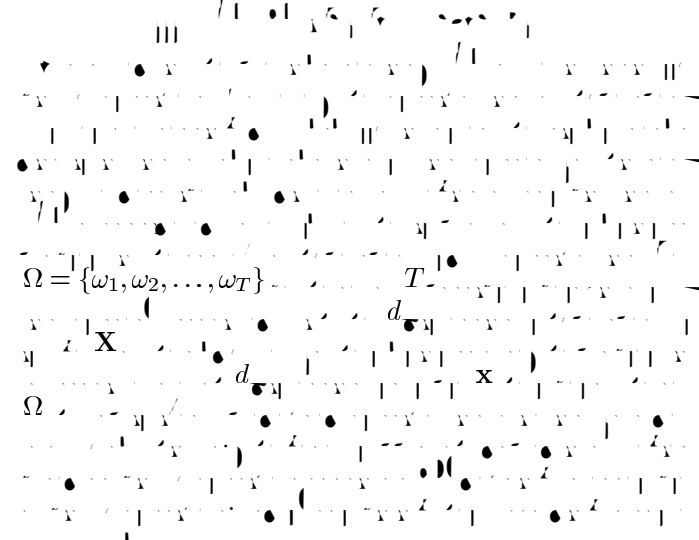
$$S_i(\mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^T f_{ij}(\mathbf{x}). \quad (2)$$

$$\omega^* = \arg \max_{i=1, \dots, T} \{S_i(\mathbf{x})\}. \quad (3)$$

B. Hierarchical Tree-Based Approach

The hierarchical tree-based approach involves building a decision tree where each internal node represents a pairwise comparison between SVMs. The root node compares all SVMs, and subsequent nodes compare the winners of previous comparisons until a single winner is identified. This approach reduces the number of comparisons from $T(T-1)/2$ to $T-1$.

Let $\mathbf{X} = \{x_1, \dots, x_T\}$ be a set of T feature vectors, M SVMs, $T(T-1)/2$



$$\begin{cases} \Omega_A = \omega_i \\ \Omega_B = \omega_j. \end{cases} \quad (1)$$

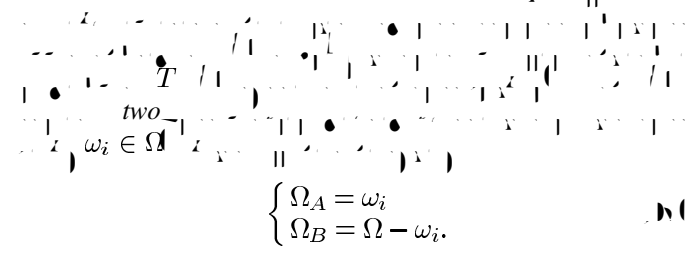
$\omega_i \in \Omega$ $S_i(\mathbf{x})$

$$S_i(\mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^T f_{ij}(\mathbf{x}). \quad (2)$$

$$\omega^* = \arg \max_{i=1, \dots, T} \{S_i(\mathbf{x})\}. \quad (3)$$

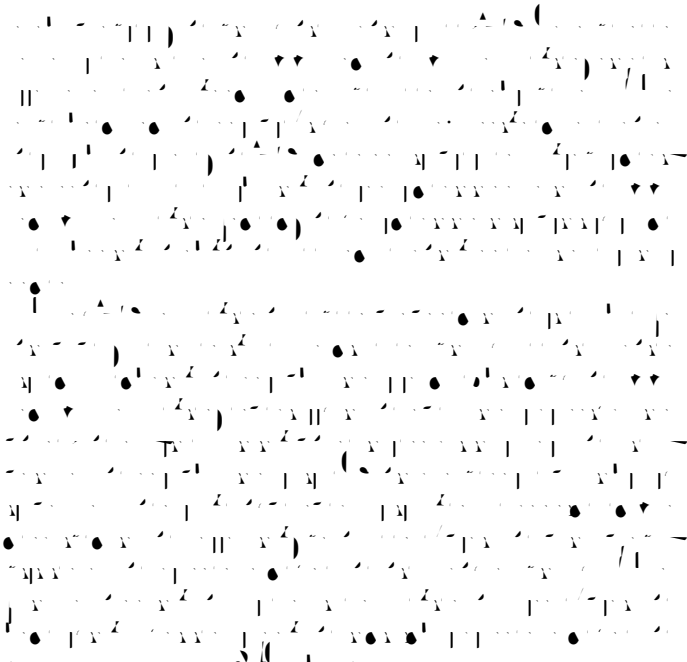
A. Parallel Approach

1) One-Against-All Strategy:

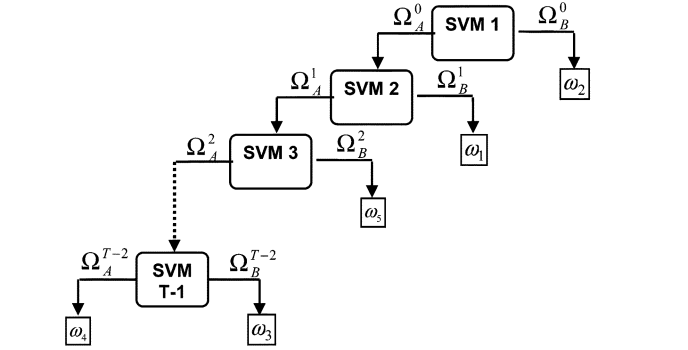


2) One-Against-One Strategy:

The one-against-one strategy involves comparing every pair of SVMs. This results in $T(T-1)/2$ pairwise comparisons. The final classification result is determined by the SVM that wins the most pairwise comparisons.

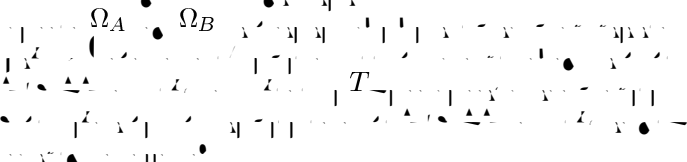


(a)



(b)

1) BHT-Balanced Branches Strategy:



Step 0: Root Node

$$k = 0$$

$$\sum_{\omega_i \in \Omega_{A,0}^k} P(\omega_i) \approx \sum_{\omega_j \in \Omega_{B,0}^k} P(\omega_j)$$

Step 1: k-Level Branching

$$q = 0, \dots, 2k - 1 (q \geq 0)$$

$$\left\{ \begin{array}{l} \Omega_{A,q}^k \\ \Omega_{A,2q+1}^{k+1} \end{array} \right\} \geq 2, \quad \sum_{\omega_i \in \Omega_{A,2q}^{k+1}} P(\omega_i) \approx \sum_{\omega_j \in \Omega_{A,2q+1}^{k+1}} P(\omega_j)$$

$$\left\{ \begin{array}{l} \Omega_{B,q}^k \\ \Omega_{B,2q+1}^{k+1} \end{array} \right\} \geq 2, \quad \sum_{\omega_i \in \Omega_{B,2q}^{k+1}} P(\omega_i) \approx \sum_{\omega_j \in \Omega_{B,2q+1}^{k+1}} P(\omega_j)$$

$$k = k + 1$$

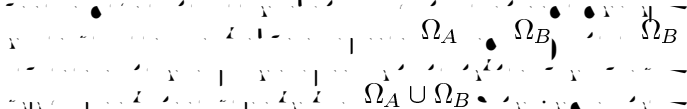
Step 2: Stop Condition

$$\exists \Omega_{A,q}^k, \Omega_{B,2q}^k, \left\{ \Omega_{A,q}^k \right\} \geq 2$$

$$\left\{ \Omega_{B,q}^k \right\} \geq 2 \quad (q = 0, \dots, 2^k - 1)$$

Step 1.

2) BHT-One Against All Strategy:



CLASS	TRAINING	TEST
ω_1 - Corn-no till	742	692
ω_2 - Corn-min till	442	392
ω_3 - Grass/Pasture	260	237
ω_4 - Grass/Trees	389	358
ω_5 - Hay-windrowed	236	253
ω_6 - Soybean-no till	487	481
ω_7 - Soybean-min till	1245	1223
ω_8 - Soybean-clean till	305	309
ω_9 - Woods	651	643
Total	4757	4588

Step 0: Root Node

$$k = 0$$

$$P(\Omega_B^k)_{\Omega_B \in \Omega} = \max_{\omega_j \in \Omega} \{P(\omega_j)\} \quad \Omega_A^k = \Omega - \Omega_B^k$$

Step 1: k-Level Branching

$$P(\Omega_B^{k+1})_{\Omega_B^{k+1} \in \Omega_A^k} = \max_{\omega_j \in \Omega_A^k} \{P(\omega_j)\}$$

$$\Omega_A^{k+1} = \Omega_A^k - \Omega_B^{k+1}$$

$$k = k + 1$$

Step 2: Stop Condition

$$\left\{ \Omega_A^k \right\} \geq 2 \quad \text{Step 1.}$$

$$T \cdot \frac{T(T-1)}{2}$$

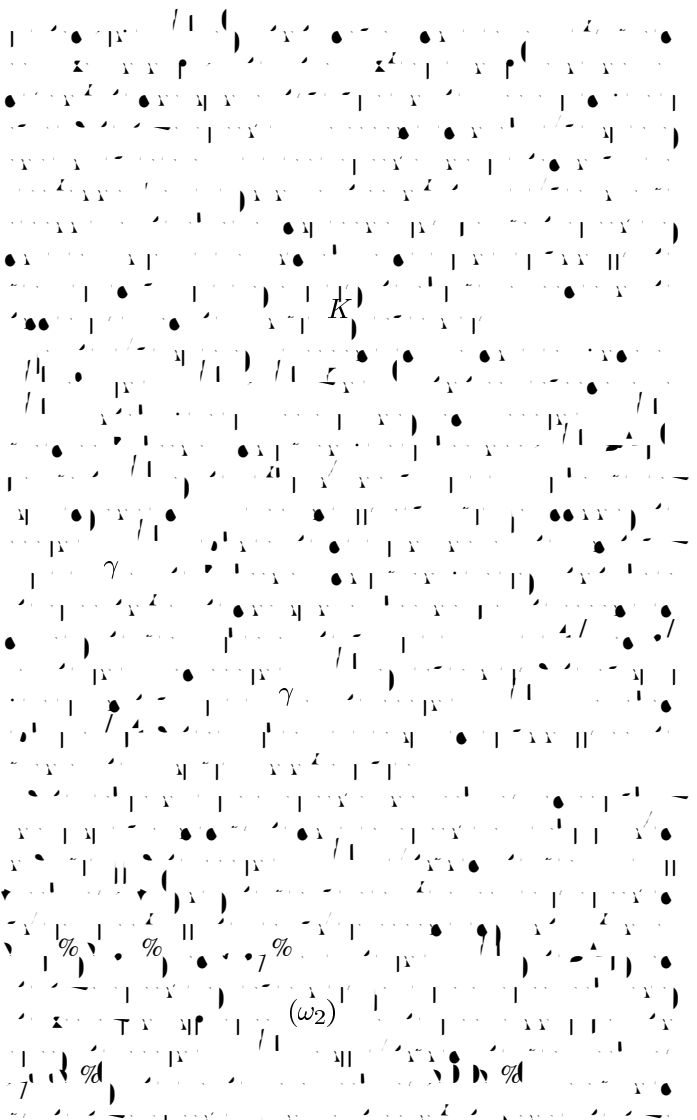
$$T - 1$$

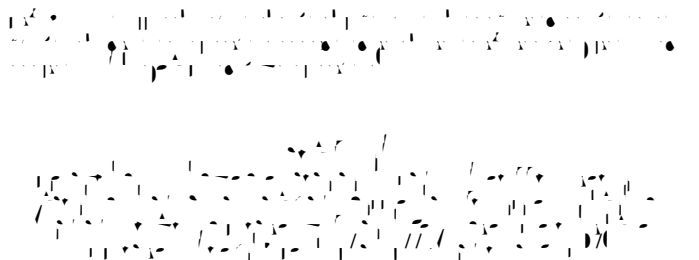
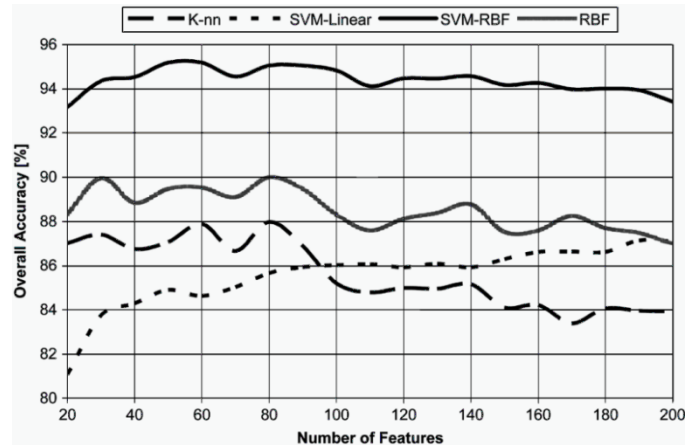
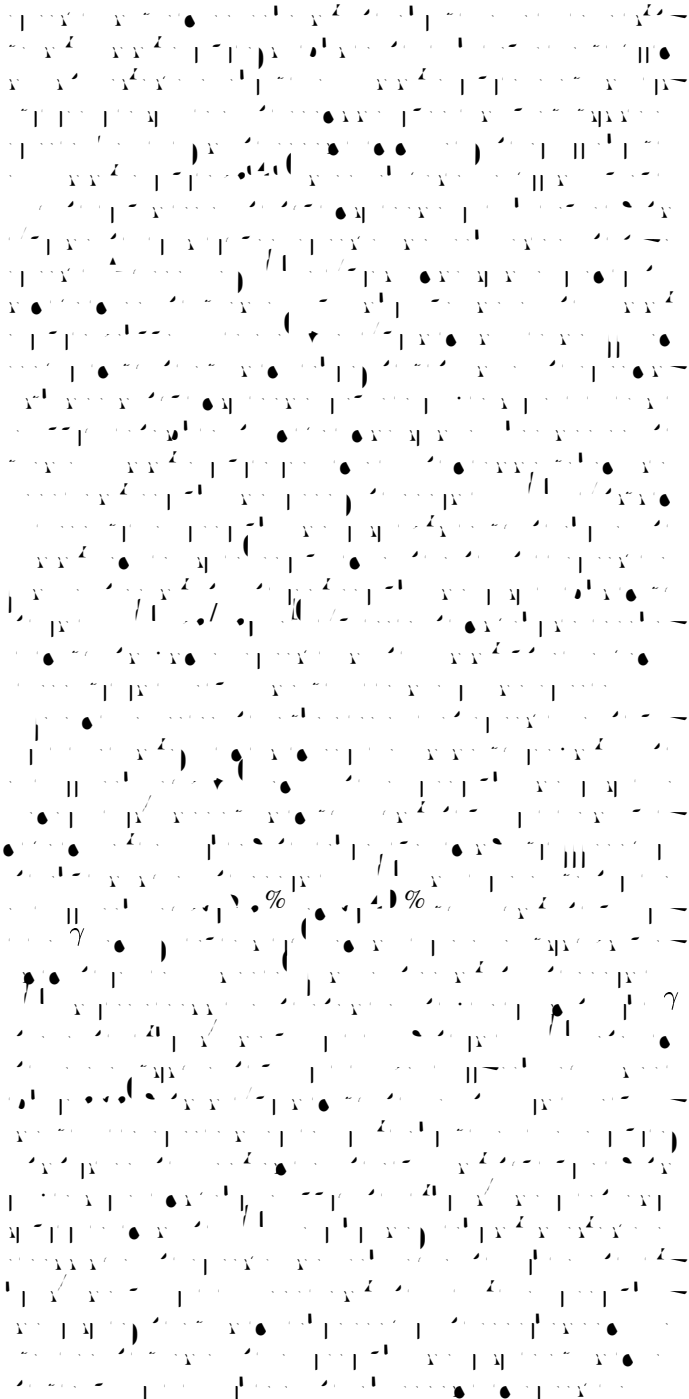
METHOD	CLASSIFICATION ACCURACY [%]										TIME [s]
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	OA	
SVM-Linear	89.02	69.13	94.51	98.60	100	75.47	83.48	83.17	99.22	87.10	40342
SVM-RBF	91.47	87.76	94.94	98.88	100	88.57	91.25	95.79	99.38	93.42	2702
K-nn classifier	96.73	61.16	86.59	80.46	99.60	98.88	90.72	65.82	74.42	83.94	2618
RBF classifier	98.44	74.11	88.47	79.83	99.21	98.04	91.98	73.72	80.06	86.99	4743

A. Dataset Description and Experiment Design

METHOD	PARAMETER RANGE	OVERALL ACCURACY [%]		MEAN TOTAL TIME [s]
		Mean	Variance	
SVM-Linear	$C \in [1, 100]$	85.38	4.94	20785
SVM-RBF	$C \in [1, 100]; \gamma = 1$	92.64	0.84	1695
SVM-RBF	$\gamma \in [0.1, 3]; C = 40$	92.51	0.50	2412
K-nn classifier	$K \in [1, 25]$	82.42	1.56	2630
RBF classifier	$N^\circ \text{ clusters} \in [20, 200]$	85.59	1.12	1505

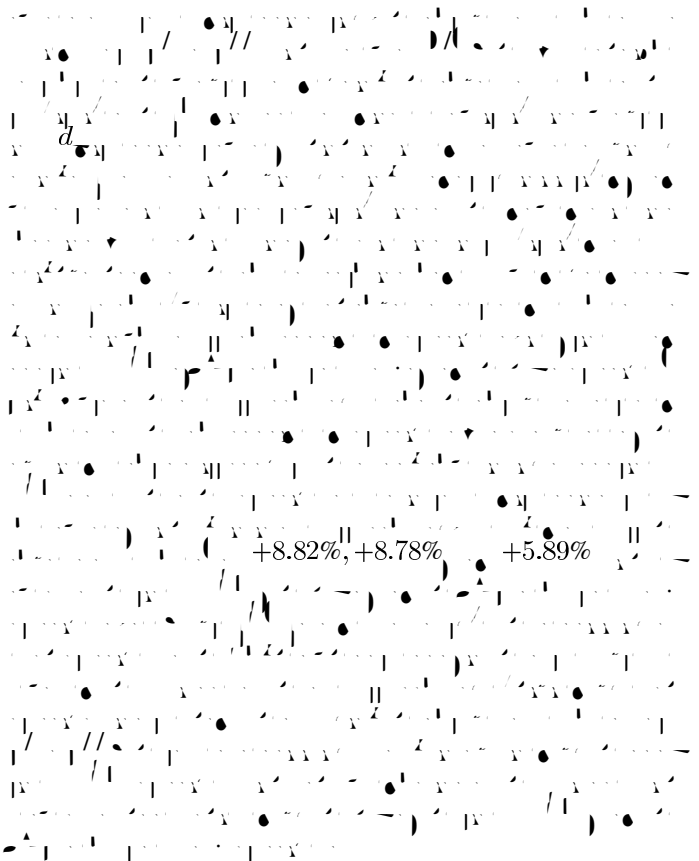
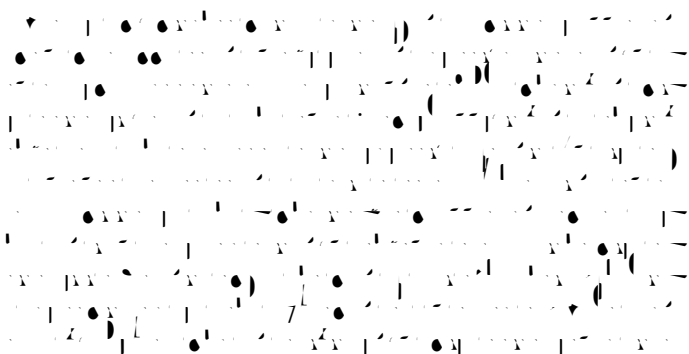
B. Results of Experiment 1: Classification in the Original Hyperdimensional Feature Space



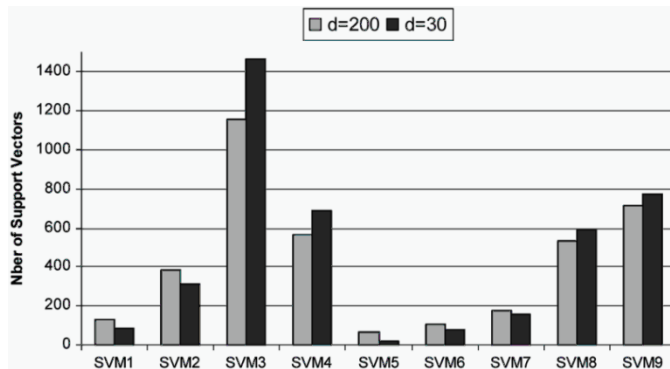


METHOD	OVERALL ACCURACY [%]	
	Mean	Variance
SA + SVM-Linear	85.56	2.04
SA + SVM-RBF	94.38	0.30
SA + K-nn classifier	85.60	2.27
SA + RBF classifier	88.49	0.81

C. Results of Experiment 2: Feature Reduction and Classification



METHOD	CLASSIFICATION ACCURACY [%]										DIFF-OA [%]
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	OA	
SA + SVM-Linear	98.29	73.79	81.11	72.35	100	99.16	89.45	57.4	86.27	83.74	-3.36
SA + SVM-RBF	99.69	92.23	93.30	91.48	99.6	99.72	97.89	88.52	91.62	94.35	0.93
SA + K-nn classifier	80.93	71.43	94.52	99.44	99.61	87.53	88.06	71.20	96.42	87.40	3.46
SA + RBF classifier	83.24	77.30	93.25	97.77	98.03	86.08	90.03	77.67	98.29	89.95	2.96



+10.69%, +7.21% +5.82%

(d = 200)
(d = 30)

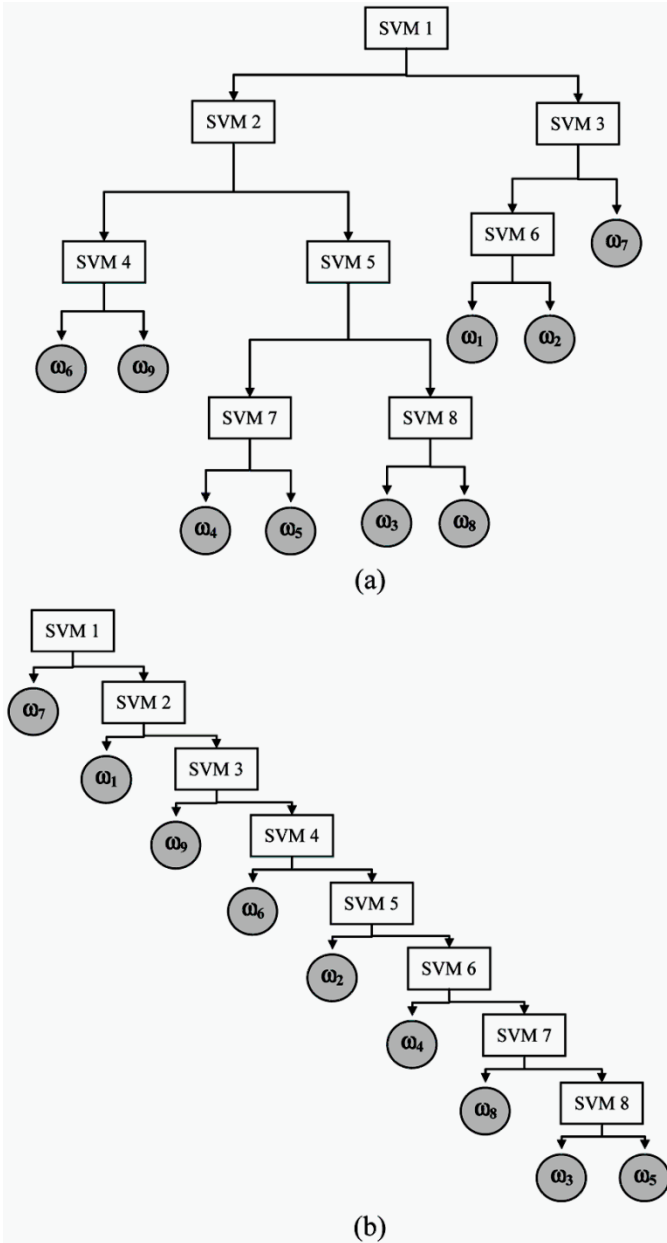
+2.72%, +1.72% +0.54%

two

11%

D. Results of Experiment 3: SVM and Multiclass Strategies

Multiclass Strategy	CLASSIFICATION ACCURACY [%]									
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	OA
OAA	91.47	87.76	94.94	98.88	100	88.57	91.25	95.79	99.38	93.42
OAO	90.32	89.54	94.51	99.72	100	88.36	93.54	94.82	99.38	93.96
BHT-BB	89.70	87.25	95.78	99.16	100	87.73	91.17	85.11	98.60	92.24
BHT-OAA	88.87	85.97	95.36	98.88	100	85.24	87.82	90.62	99.07	91.24



Multiclass Strategy	TIME [S]		NUMBER OF SVMs	number of Support Vectors		
	Train	Test		Min	Max	Average
OAA	2361	341	9	62	1159	424
OAO	212	554	36	9	569	130
BHT-BB	311	125	8	31	977	334
BHT-OAA	410	155	8	45	1270	333

Multiclass Strategy	OVERALL ACCURACY [%]							
	SVM1	SVM2	SVM3	SVM4	SVM5	SVM6	SVM7	SVM8
BHT-BB	95,66	99,25	94,1	100	99,57	98,62	100	98,53
BHT-OAA	93,77	97,74	99,7	99,16	98,9	99,57	99,12	100

IEEE Trans. Inform. Theory

IEEE Trans. Pattern Anal. Machine Intell.

IEEE Trans. Geosci. Remote Sensing

IEEE Trans. Geosci. Remote Sensing

IEEE Trans. Geosci. Remote Sensing

light

J. R. Statist. Soc.

Signal Process. Mag.

Remote Sensing Digital Image Analysis

IEEE Trans. Geosci. Remote Sensing

Pattern Recognition and Signal Processing

Pattern Recognit. Lett.

IEEE Trans. Geosci. Remote Sensing

IEEE Trans. Pattern Anal. Machine Intell.

IEEE Trans. Geosci. Remote Sensing

IEEE Trans. Geosci. Remote Sensing

Remote Sens. Environ.

IEEE Trans. Geosci. Remote Sensing

IEEE Trans. Geosci. Remote Sensing

Proc. IGARSS

Proc. SPIE

Int. J. Remote Sens.

Proc. SPIE

Summaries 8th JPL Airborne Earth Science Workshop

Proc. IGARSS

Proc. IGARSS

Statistical Learning Theory

Proc. 5th Annu. ACM Workshop Computational Learning Theory

Data Mining Knowl. Discov.

IEEE Trans. Med. Imag.

IEEE Trans. Neural Networks

IEEE Trans. Pattern Anal. Machine Intell.

IEEE Trans. Signal Processing

IEEE Trans. Neural Networks

Mach. Learn.

Neural. Comput.

IEEE Trans. Syst., Man, Cybern. C

A Course in the Geometry of n-Dimensions

Introduction to Statistical Pattern Recognition

Proc. Int. Conf. Pattern Recognition

Advances in Kernel Methods: Support Vector Learning

IEEE Trans. Geosci. Electron.

IEEE Trans. Geosci. Remote Sensing

Proc. 3rd Int. Workshop on Multiple Classifier Systems—MCS 2002

IEEE Trans. Geosci. Remote Sensing

Sensing

IEEE Trans. Geosci Remote. Sensing



Farid Melgani



Lorenzo Bruzzone

Recognition of IEEE Transactions on Geoscience and Remote Sensing Best Reviewers