Chapter 1

Introduction

Remote sensing is the science of observing the earth from a distance, mostly using airborne or satellite sensors, to study the occurrence of different phenomenon on it [144]. Many substances on the earth can be distinguished by their appearance in a color image however, a three band color information may not be sufficient to distinguish the internal composition of the substance. Hyperspectral images have the capability to distinguish the substances based on their internal composition [40]. Therefore, the analysis of hyperspectral images is helpful in land-cover/land-use mapping, agriculture, soil-mapping, forestry, and many more applications [77]. However, the analysis of hyperspectral images is not free from challenges which includes their huge data volume and limited availability of training samples [19]. Moreover, the classic analysis of hyperspectral images are only dependent on the spectral values (reflected light intensity received by the sensor) and do not consider their spatial information [56, 63, 80]. In this thesis, we are proposing some advance algorithms for analysis of hyperspectral images using the integrated spectral and spatial information.

This chapter presents an overview of the thesis. First, the hyperspectral image analysis and its related challenges are discussed. Then, the background literature is reviewed after which the objective and main contributions of the thesis are discussed. Finally, the organization of the thesis is discussed at the end of the chapter.

1.1 Hyperspectral remote sensing image analysis

Hyperspectral remote sensing deals with acquiring digital imagery of earth surface in hundreds of narrow contiguous spectral bands [40]. This gives a rich information to recognize the materials on the earth surface. This section describes the principle of hyperspectral remote sensing images, its applications and challenges.

1.1.1 Principles of hyperspectral remote sensing images

In remote sensing, we acquire images of earth from a distance. This task is accomplished using an airborne or a satellite sensor [144]. The images can be analog or digital. Analog images are acquired by photographic sensors on paper based or transparent media. Such images can be enlarged to any extent without blurring the image. However, a computer can understand it only after converting it into a matrix where each cell stores the light intensity of the respective part of the image. The cell of an image is also called pic-cell or *pixel*. In contrast, a digital image is acquired by an electro-optical sensor that stores the image directly as a rectangular array of cells (*pixels*), because of which it is easily readable by the computers. The radiant energy received from a portion of earth surface is stored as a numeric value (a digital number) in a pixel of digital image.

In remote sensing, there are mainly three modes of digital imagery depending on the number of discrete bands considered while acquisition of the image. In a beam of electromagnetic radiation, each discrete wavelength carries different amount of energy. Therefore, while acquisition of an image the sensor stores radiant energy carried by one or more discrete wavelength bands. Based on the number of bands considered during acquisition, the modes of digital imagery can be categorized as panchromatic, multi-spectral and hyperspectral. The panchromatic images consider a single band that corresponds to visible range of electromagnetic spectrum. They display shades of gray and are generally acquired in high resolution. A color image is acquired in three bands each in range of blue, green and red. A multi-spectral image is acquired in multiple spectral bands (usually 4 to 10) ranging from visible through infra red (IR) region. When more than 10 spectral bands are considered in image acquisition then the acquired image is called hyperspectral image (HSI) [40].

Usually, HSIs are acquired in hundreds of contiguous spectral bands with a

Table 1.1: Specification of some hyperspectral sensors.

Sensor	Platform	Spectral channels	Spectral range	Spectral resolution	Spatial resolution
Hyperion	Satellite	220	400-2500nm	10nm	30m
MODIS	Satellite	36	620-965nm 3660-14280nm	40nm	250m 500m 1000m
AVIRIS	Aerial	224	400-2500nm	10nm	20m
ROSIS-3	Aerial	115	430-850nm	4nm	1.3m
CASI	Aerial	288	380-1050	2.5nm	1-10m
HYDICE	Aerial	210	413-2504nm	10nm	1-10m
HYSI	Satellite	64	421-964nm	20nm	80m
PRISMA	Satellite	250	400-2500nm	12nm	30m

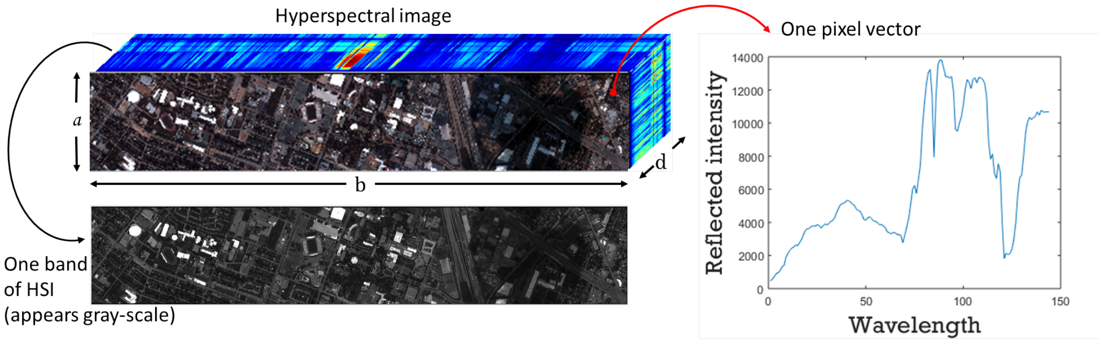

Figure 1-1: Representation of a hyperspectral image. For each pixel in the scene, reflected light intensity is recorded in multiple spectral bands. Images acquired in one band appears gray-scale.

fine spectral resolution. The spectral resolution is defined as the distance between the considered contiguous spectral bands [77]. The portion of earth covered by a pixel determines the spatial resolution of the image [117]. The spectral and spatial resolution, number of spectral bands and coverage of electromagnetic spectrum may differ from sensor to sensor. Table 1.1 shows a list of well known hyperspectral sensors and their specifications. One can see from the table that the satellite based sensors usually have lower spatial resolution than the areal sensors although they have similar spectral coverage. The Aerial platform corresponds to airborne sensors. The images acquired by airborne sensors are usually of higher spatial resolution than those acquired by a satellite. Hyperspectral images acquired by these sensors are a rich source of information. Fig. 1-1 demonstrate the concept of hyperspectral imagery. These images are stored as a three dimensional matrix $(a \times b \times d)$ where the first two dimensions (a and b) are the spatial dimension whereas the third (i.e. d) is the spectral dimension [100, 189]. One can observe from the Fig. 1-1 that each pixel of HSI is represented as a vector having reflected light intensity from earth's surface in multiple wavelength. The set of reflected intensity from the earth's surface in multiple wavelength is known as its spectral signature [39]. Since different substances on the earth surface manifest different spectral signature, the hyperspectral imaging technology can easily distinguish the abundance of material and their composition on the surface of earth. Therefore, they are useful in several remote sensing applications [20] such as:

- *Mining/ Geology*: Analysis of an HSI helps in mapping the abundance of minerals and finding concentration of heavy metals in stream sediments [129, 141, 242].
- *Forestry*: Along with land-cover classification, analysis of an HSI is useful in identifying species, vegetation stress, detailed stand mapping, foliar chemistry and so on [17, 122, 184].
- *Aggriculture*: Along with crop mapping, an HSI is helpful in soil mapping, crop differentiation and monitoring crop stress [1, 8, 54, 61, 225].
- *Monitoring and environmental management*: HSI is being used to effectively monitor the land-cover and changes in the land and coastal-ocean ecosystem. This helps in monitoring urban growth and generating alarm for probable natural disasters [7, 35, 47, 51, 78, 94, 123, 126, 198, 214].
- *Military applications*: Due to its effectiveness in detecting the objects that are not visible to the necked eye, analysis of HSI is a powerful tool for military surveillance [6, 24, 149, 151, 218, 238].

HSI is definitely a rich source of information but its analysis is not straight forward. The techniques developed for analysis of gray-scale, color or multi-spectral images are not suitable for analyzing hyperspectral images [96]. Therefore, in last decade several developments have been proposed in literature for analysis of hyperspectral images which shows the growing interest of researchers in this field. The analysis of HSI involves *unmixing* [18, 104, 127, 145, 167, 173], *segmentation* [139, 168, 221, 223], *change detection* [29, 146, 155, 156] and *pixel-wise classification* [32, 46, 57, 63–65, 82, 162, 224, 249]. This thesis is devoted to pixel-wise classification of HSI. Several literature methods have attempted to counter the challenges in pixel-wise classification of HSI. However, still the pixel-wise classification of HSI is not free from challenges. The next subsection describes the pixel-wise classification of HSI along with some of its challenges.

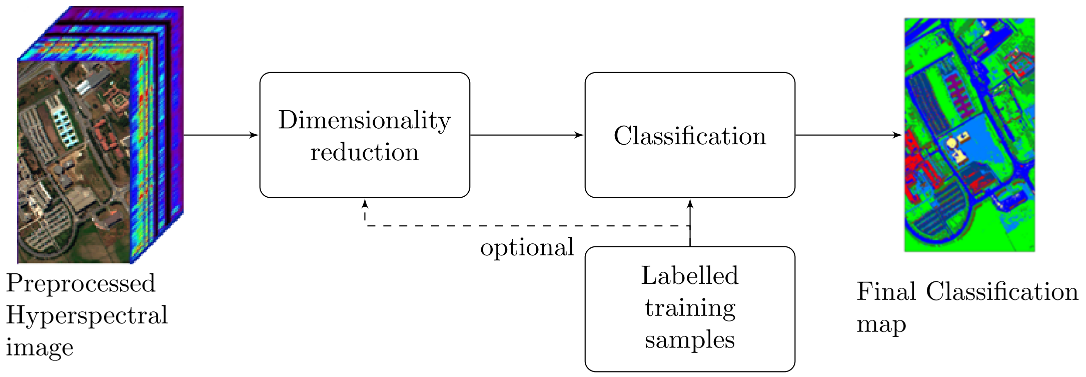

Figure 1-2: General framework for pixel-wise classification of Hyperspectral images.

1.1.2 Pixel-wise classification of HSI

Pixel-wise classification of hyperspectral image involves the assignment of a class label to each pixel of the image. The pixel-wise classification of HSI can be unsupervised or supervised. In unsupervised classification, the pixels are grouped together considering some similarity measure. No labelled samples are required for this [168, 221]. In contrast, supervised classification uses labelled samples for modeling a classifier that can recognize class label for each pixel of the given image [32, 64, 162]. In pixel-wise classification of HSI H, having d spectral bands ($\{G_1, G_2, G_3, ..., G_d\}$), each of its pixel is a 'd' dimensional vector that is considered as a pattern and the task is to assign a class label to it from the list of the predefined classes $C_1, C_2, ..., C_c$. For this, some samples from each class are labelled manually either by field survey or image interpretation in lab. These labelled samples are used as supervised information based on which a classifier model is constructed to recognize the class label for all the unlabelled pixels of the scene in order to generate a classification map. All the pixels of same class are assigned same color in the classification map. The class labels can be wheat, maize, paddy, etc. for an agriculture field [1, 54]. It can be trees, roads, roof-top, shadows, etc. for an urban scene [57, 82] and it can be name of minerals for mineral mapping application [141, 242].

The classification of HSI is a step by step process. The acquired HSI is preprocessed for atmospheric correction, geometric correction and radiometric calibration. After preprocessing the HSI data is ready for analysis. The classification process of HSI is demonstrated in Fig. 1-2. Since the HSI has a large dimension that limits the statistical estimation and effects the classification process, the dimension of HSI is reduced using any supervised or unsupervised method. After dimensionality reduction the unlabelled pixels are classified using a classifier model designed based on available labelled samples. This classification leads to classification map which identifies pixels of different classes in distinct colors [134, 162].

For the supervised classification of HSI several classifier models are suggested in the literature. The primitive approach to pixel-wise classification analyzed the distance of the pixel vectors from the mean vector. The distance can be based on euclidean distance [136], Jeffries-Matusita distance [133], spectral angle [203] information divergence [39], transformed divergence [101], etc. These distance based classification methods work fine when the inter-class variability is low. Otherwise, such algorithms provide less accurate results [136].

Gaussian maximum likelihood classifier has been widely used in literature for classification of HSI [119, 130, 136, 192, 194]. it classifies a given pixel x to a class C_j for which the posterior probability $P(C_j/x)$ is maximum. This method has a drawback that it assumes the shape of probability distribution function to be Gaussian. In case the assumption is not true, the classifier produces inaccurate results [170]. Moreover, the large dimension and limited availability of training samples adversely effects the estimation of statistical parameters.

Another approach to classification which is free from statistical distribution of data is based on neural networks. Researchers all over the world showed interest in classification of HSI based on neural networks [170]. However, it has a limitation of algorithmic and training complexity [162].

In recent years, support vector machine (SVM) has gained popularity [10, 46, 63, 162, 163]. It uses kernel methods to transform data into another feature space where the data appears linearly separable. Because of this feature SVM is quite suitable for high dimensional data and hence for classification of HSI. In [162] Melgami and Bruzzone have shown that the SVM is more effective than neural networks and K-nearest neighbor classifier.

1.1.3 Challenges in classification of HSI

The classification of HSI has three important challenges. First is the dimensionality reduction. The dimensionality reduction algorithm should preserve the most of the information content of HSI. Second is the acquisition of labelled samples. Since the labelling of pixels is costly and time consuming task, the pixels that are most informative to discriminate well among classes should be labelled. The third challenge is integration of spectral and spatial information. Since integration of spectral and spatial information involve different image processing techniques, many of which are parameter dependent, constructing a spectral-spatial profile that is of low dimension and represents maximum information is a challenging task. All these challenges are explained in the following.

Huge data volume

With rich spectral information covering hundreds of wavelengths an HSI is having huge data volume. This huge volume requires large storage space and high computational time during analysis. The major problem with huge data volume is *curse of dimensionality* [12]. This problem states that as the dimension of the data increases, more number of training samples are required for proper statistical estimations [88]. However, the hyperspectral remote sensing images have limited number of training samples. This creates an adverse effect on the classification performance of any classifier model. This is known as *Hughes phenomenon*, where the classification performance improves with increase in data dimension but it starts deteriorating as the dimension, in terms of spectral channel, keeps increasing [115]. Fig. 1-3 visualizes this behavior of data. Furthermore, the high dimensionality of hyperspectral data makes it necessary to seek new analytic methods to avoid the vast increase in the computational time. A challenging yet effective way to deal with these complexities, arising due to high dimension, is to reduce the dimension [103].

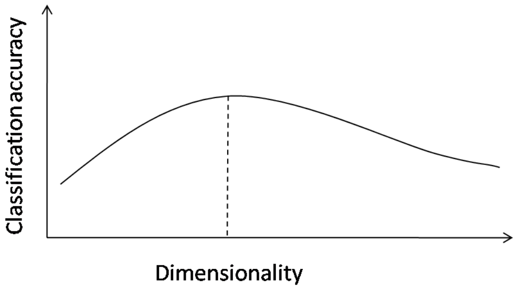

Figure 1-3: Hughes Phenomenon: Y axis represents classification accuracy and X axis represents dimensionality. If the number of samples are constant, the classification accuracy increases till a few features and after that it actually decreases with the increase in dimensionality.

Limited availability of training samples

Fukunaga [88] convincingly explained that the dimension of the data and the required number of training samples for different classifier have a relation. The increase in dimension of data for linear, quadratic and non-parametric classifiers is related to the required number of training samples linearly, to the square and exponentially, respectively. In words of remote sensing, as the number of spectral channels increase, supervised classification needs more number of training samples for accurate classification. Labelling of samples for remote sensing HSI is a lengthy task. It can be accomplished through ground survey and/or photo interpretation in lab which requires huge amount of time and effort. Furthermore, some of the labelled pixels may provide redundant information. In a situation where labelling pixels is so costly, one should be judicious while selecting a pixel to label. In other words, it is a challenge to select the most informative pixels in the image for labelling. This problem is addressed by active learning (AL) techniques which selects most informative pixels on the basis of uncertainty, diversity and other criteria [229].

Incorporation of spatial information

Conventional pixel-wise classification of HSI considers the spectral measurements only and no spatial information [134, 240]. In practice, pixels are spatially related due to the homogeneous distribution of land cover. Information captured in neighboring locations may provide useful supplementary knowledge for analysis of a pixel. Therefore, spectral information with the support of spatial information can effectively reduce the uncertainty of class judgment. Moreover, the new generation of spectrometers are producing high resolution hyperspectral images. With the rich spectral information of earth's surface, it also increases intra-class variability [25]. Because of this the spectral signature varies in the same class which makes the analysis of HSI less effective. Therefore there is a need to integrate the spectral and spatial information for analysis of HSI [14, 222]. It has been observed that the integration of spectral and spatial information can improve classification accuracy and reduce salt and pepper effect of the classification map [62, 95, 128]. Yet, deriving spatial information out of hyperspectral remote sensing image for each pixel, that could help in increasing class separability, is a challenging task. Moreover, the integration of spectral and spatial information involves image processing operations, many of which are dependent on parameters. Many a times the integrated spectral-spatial information turns into a large dimensional data. Therefore, constructing a low dimensional spectral-spatial profile representing maximum spatial information is a challenging task.

1.2 Background literature

Hyperspectral images poses rich spectral information but along with that it has large dimension, huge data volume and limited availability of training samples. Researchers worldwide are showing their interest in mitigating these challenges for the effective analysis of HSI. Along with the computational complexity, the classification of HSI has to face *Hughes phenomenon* [115]. So, reducing the dimension of data by maintaining sufficient information content is an effective technique. Even after reducing the dimension, supervised classification suffers due to limited availability of labelled training samples. To classify an HSI with limited number of training samples, two approaches exist in the literature. One is semi-supervised learning and another is active learning (AL). Semi-supervised learning incorporates both the labelled and unlabelled data into the training phase of a classifier to obtain better decision boundaries [33, 37, 148, 154, 204, 235]. In contrast, AL is a paradigm to reduce the labeling effort and optimize the performance of a classifier by including only most informative patterns (which have highest training information for supervised learning) into the training set. Another issue in HSI classification is the increase in the variability of spectral signature within the class with the increase in spatial resolution. Since new sensors are acquiring high resolution HSIs, considering only spectral values while classification may not be sufficient for accurate classification. So, integration of spectral and spatial information is important for analysis of HSI. However, integration of spatial information with spectral information is a challenging task. Therefore, in the following subsections the literature related to dimensionality reduction, active learning and approaches on integrating spectral and spatial information is presented. The literature review is focusing on principal component analysis, SVM based active learning techniques and mathematical morphology based integration of spectral-spatial information, which are the techniques focused in the thesis.

1.2.1 Dimensionality reduction

Hyperspectral data poses high dimensional feature representation which provides rich information along with some mathematical challenges. Because of high dimension and small number of available labelled samples statistical estimations are less accurate. This is the curse of dimensionality [12], that states more number of training samples are required with the increase in dimension of data for proper statistical estimation. To avoid the problems arising due to large volume, dimensionality reduction has been suggested in literature. Different types of dimensionality reduction techniques are discussed in [103]. Broadly, feature selection (selection of subset of features from available) and feature extraction (transforming available feature space to a new feature space) are two ways of dimensionality reduction in the literature.

Feature selection

Feature selection is the process of selecting subset of most informative and discriminative features from the set of features. In case of hyperspectral image, feature selection is also known as band selection. The feature selection methods are classified in following groups.

- Filter: This group of methods use to select the features by ranking them based on their information content and do not consider any classifier while selecting features. The ranking of the features can be based on information gain [3, 137, 212], mutual information [108, 132, 185, 211, 219], correlation [125, 241], regression coefficient [109], rough set theory [9, 42, 180, 234], clustering [118] etc. These methods are simple, scalable and shows good empirical success.
- Wrapper: The feature selection methods in this category decides the discriminative power of features using a predictor (classifier). These methods search all the probable subset of features and assess their prediction performance for guiding the search and halting it [153]. The exhaustive search of all possible subsets is NP-Hard problem. However, one can exploit different techniques like best-fit, branch and bound, simulated annealing, genetic algorithms [252], self-adaptive differential evolution [97] *etc.*
- Embedded: In these methods the features are selected while learning process of predictor. They reach a solution faster and uses the available data in a better way [5, 243, 251].
- Hybrid: The filter methods are simple but less accurate whereas wrappers are more accurate but need more computational power. The hybrid methods

are an attempt to combine the benefits of both the methods together [60, 161, 169].

In hyperspectral image analysis each pixel is a pattern represented as a vector having reflected light intensity in different spectral bands. These spectral bands are treated as features in HSI classification. Several methods exist in the literature for feature selection (band selection) of HSI . Some of them use labelled samples to find a group of bands achieving the largest class separability [9, 97, 180, 252] and others try to do the same in an unsupervised way. The unsupervised methods are based on uniform spectral similarity [74], information measure [84, 102], clustering [118, 160] etc. In current trend, feature selection methods based on stochastic and evolutionary optimization techniques (e.g., ant colony optimization [42, 233], Genetic Algorithms (GA) [169, 252] and Particle Swarm Optimization (PSO) [234]) are popular.

Feature extraction

Feature extraction is the process of transforming data from one feature space to another feature space where recognition of informative features is relatively easy. These can also be supervised and unsupervised. Some supervised feature extraction techniques are Discriminant Analysis Feature Extraction (DAFE) [88], Decision Boundary Feature Extraction (DBFE) [135], Nonparamatric Weighted Feature Extraction (NWFE) [131], Independent Component Analysis (ICA) [49]. Some unsupervised methods are PCA (Principle Component Analysis) [120], Kernel PCA [201], etc. These have been applied throughout the literature of hyperspectral image classification[110, 124]. PCA is a popular choice in literature for dimensionality reduction of HSI because of its simplicity and unsupervised nature. In this thesis, PCA is being used for reducing the dimension of HSI in all the experiments. So, PCA is explained in the following.

PCA is an orthogonal transformation technique widely used in feature extraction and data compression [182]. It transforms a set of patterns from a d-dimensional original feature space into a d-dimensional new feature space where the transformed features are called principal components (PCs). The transformation is defined in such a way that the first PC has the largest possible variance of the patterns, and each succeeding component in turn has the highest variance possible under the constraint to be orthogonal to the preceding components. Fig. 1-4 illustrates the working of PCA using a synthetic data where the first PC is in the direction of maximum variance. The variance of the data covered by each PC determines how informative the PC is. Thus, PCA orders the PCs according to the variance of the patterns. It is used to reduce the dimensionality of the data by keeping first few PCs.

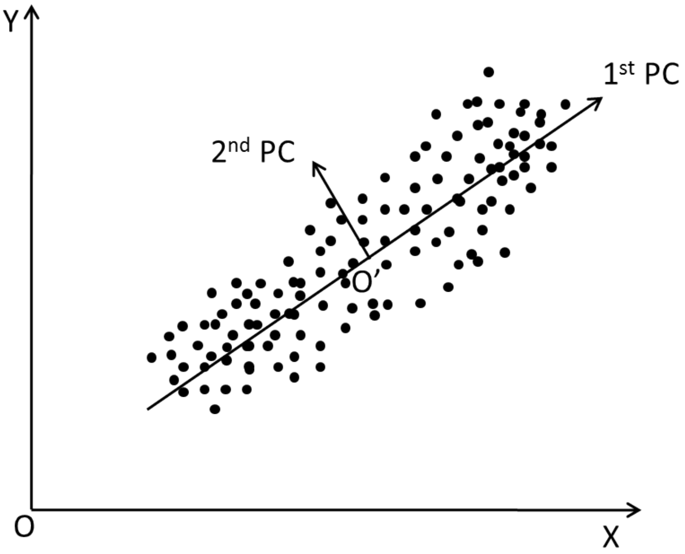

Figure 1-4: The concept of principal component analysis. The data is transformed into a new dimension where the direction having maximum variance becomes first PC and the next high variance becomes second PC and so on.

The aim of PCA is to transform data into a new feature space where each dimension (PC) is in direction of maximum variance being orthogonal to other PCs. To accomplish this we start with the $n \times d$ data matrix (X) having n patterns ($n = a \times b$ for an $a \times b \times d$ HSI data) and 'd' features (spectral bands $\{G_1, G_2, G_3, ..., G_d\}$) ignoring the class labels. Please note that no supervised information is required for PCA. Following steps are involved in the PCA of data matrix X.

1. *Normalization*: Each feature (G_i) of the data matrix X is normalized. This is done to restrict the domination of one feature over another. For example, the feature having values in the range of 0 to 400 may dominate the feature having value in range of 0 to 1. This may produce biased results. Normalization brings all the features in similar range. For normalization, one has to subtract the mean of the feature from each value of the feature and divide by standard deviation. Equation 1.1 calculates a normalized feature vector $Norm\_G_i$ for the ith feature vector G_i where $G_i(k)$ is the kth value, $\mu(G_i)$ is the mean and $\sigma(G_i)$ is the standard deviation of the ith feature in data matrix X.

$$Norm\_G_i(k) = \frac{G_i(k) - \mu(G_i)}{\sigma(G_i)} \tag{1.1}$$

By normalizing all the d features of data matrix X we get a $n \times d$ normalized matrix referred hereafter as $Norm\_X$.

2. *Computation of covariance matrix*: In this step, one has to find the covariance between each pair of features in the normalized data matrix $Norm\_X$. The covariance value allows us to understand the relationship between the features. If the covariance of two features is positive, it means that they increase or decrease together, whereas a negative covariance value signifies that the features are inversely correlated and hence when one increases, the other decreases. A zero covariance value means there is no relation between features. In this step of PCA, a $d \times d$ matrix Cov (symmetric in nature) is computed that has entries of covariance values computed for each pair of features in $n \times d$ matrix $Norm\_X$.

$$Cov(i,j) = \frac{\sum_{k=1}^{n} ((G_i(k) - \mu(G_i))(G_j(k) - \mu(G_j)))}{n} \tag{1.2}$$

where $k \in \{1, 2, \ldots, n\}$ and $i, j \in \{1, 2, ..., d\}$

3. *Computation of principal component axis:* Once we have the covariance matrix, the next step is to compute the eigenvectors and eigenvalues from it. The eigenvectors V_i represent the principal component axis and the eigenvalues λ_i is a measure of variance covered by the vector V_i. The eigenvectors and eigenvalues are arranged in descending order of eigenvalues. Since the eigenvalues represent the variance covered by the corresponding eigenvector, this arrangement brings the most informative eigenvector at the first place, the next informative eigenvector at second and so on. The number of eigenvectors or principal component axis are same as in the original data matrix (i.e. d). For dimensionality reduction l eigenvectors which represent cumulative variance of almost 99% are selected. Therefore the dimension of eigenvector matrix becomes $d \times l$.

4. *Transformation of data into new dimension:* During all these phases the normalized data has remain as it is in $Norm\_X$. To reduce the dimension of the original data, it is transformed into the new principal component axis represented by the eigenvectors. The data in new dimension (New_X) is computed as follows.

$$New\_X = Norm\_X * V \tag{1.3}$$

The original data $Norm\_X$ was of dimension $n \times d$ and V has a dimension of $d \times l$, so the dimension of New_X is $n \times l$. New_X has lower dimension but preserves most of the information content of original data X.

After completion of PCA the original d dimensional data X is transformed into l dimensional feature space ($\{PC_1, PC_2, ..., PC_l\}$) which accumulates most of the cumulative variance of the data. Therefore, PCA is a useful technique in data dimensionality reduction and a popular choice among researchers [14, 56, 63, 80].

1.2.2 Active learning for classification of HSI having limited labelled samples

Analysis of hyperspectral images have to deal with high dimensionality and limited number of training samples. The dimensionality reduction procedure may decrease the spectral dimension of the HSI using a feature selection or feature extraction method. However, the classification results still rely on the quality of training samples selected for modeling the classifier. Usually the training samples selected from a remote sensing image is sparse, redundant and costly to collect. Due to the usually complex statistical distributions of the patterns belonging to different classes, informative labelled samples are essential to train a classifier. AL is a paradigm to reduce the labeling effort and optimize the performance of a classifier by including only most informative patterns (which have highest information for designing better classifier model) into the training set. AL techniques are usually based on iterative algorithms. In each iteration, one or a batch of most informative unlabelled patterns are chosen for manual labeling and the classification model is retrained with the additional labelled samples. An overall framework of AL is presented in Fig. 1-5. Initially, a few samples from each class

Algorithm 1 Basic active learning algorithm

```
Input:(a) Non-empty set of unlabelled samples $U$.
(b) Empty set of labelled samples $L$
(c) An expert $EPT$ for labelling the samples.
Output: Non-empty set of labelled samples $L$.
Select few samples (e.g. 3 samples from each class) randomly from $U$, get it
  labelled by $EPT$ and store in $L$.
repeat
    Update classifier model using labelled samples in $L$.
    Select informative samples from $U$ using a suitable query function.
    Get class labels for selected samples by $EPT$.
    Append the selected samples along with their assigned class labels to $L$.
    Remove the selected samples from $U$.
until Stopping criterion is met.
```

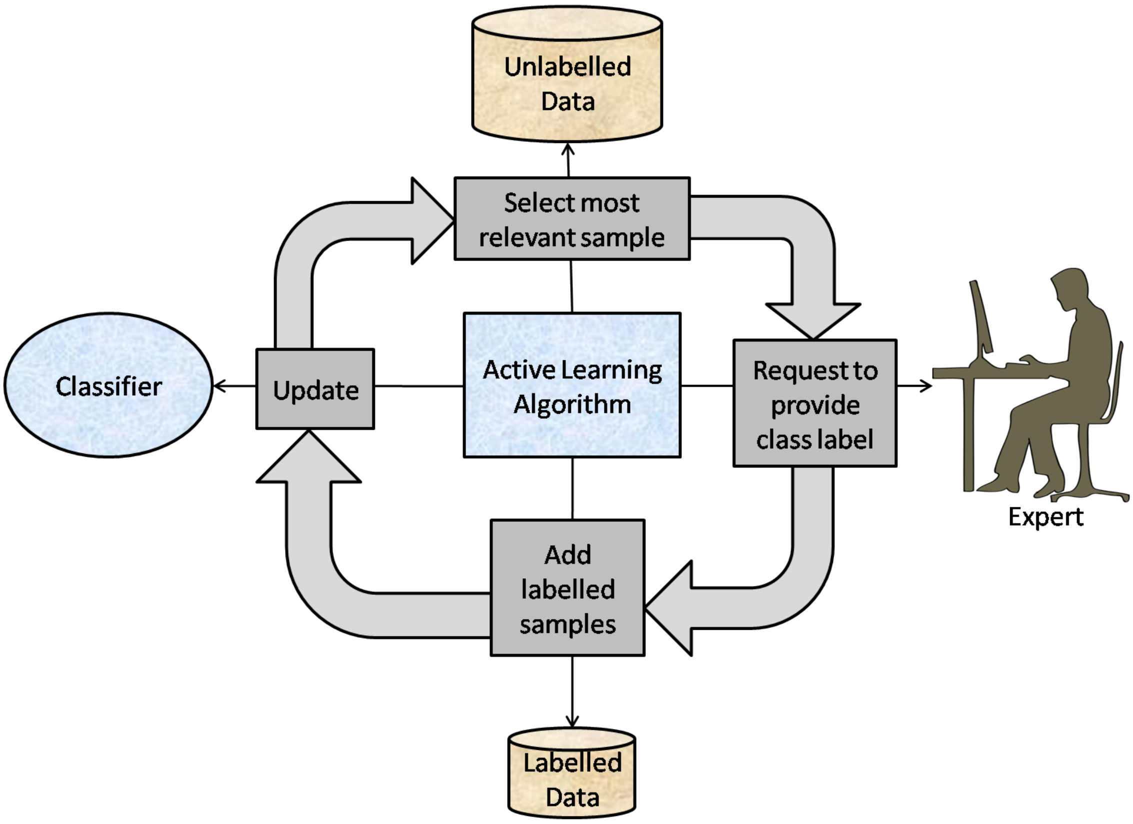
Unlabelled Data
Select most relevant sample
Classifier
Update
Active Learning Algorithm
Request to provide class label
Expert
Add labelled samples
Labelled Data

Figure 1-5: A general framework for active learning.

are randomly selected, assigned their class label and stored in labelled pool. Rest of the samples are stored in an unlabelled pool. Please note that each pixel of an HSI is treated as a sample or pattern in classification problem. After creating labelled and unlabelled pools, the labelled samples are used to train a classifier model. Next, a single or batch of samples which are most informative are selected from the unlabelled pool. Selecting informative samples from unlabelled pool is the most important part of AL procedure. The selected samples are assigned their class labels manually by an expert. While experimenting with data sets a software module can be written for labelling the selected samples. The batch of samples with their assigned class labels are appended to the labelled pool. The process of selecting informative samples is repeated until the stopping criteria is met. At the end of the iterative procedure we have informative labelled samples in the labelled pool and an updated classifier model. The basic algorithm for AL procedure is shown in Algorithm 1.

The most important component of an AL model is the selection of informative samples. This is the component that presents its own interpretation of data as well as points out the most suitable samples for including into the training set to increase the predictability of a classifier. In the literature, different AL methods are suggested based on their approach to select informative samples from unlabelled pool. These methods can be categorized in three groups based on criteria they adopt. The possible criteria are uncertainty, diversity and cluster assumption. The uncertainty criterion aims at recognizing the sample from unlabelled pool, whose class assignment is most ambiguous. The diversity criterion focuses on selecting non-redundant samples. The cluster assumption criterion aims at recognizing the samples in low density region of feature space assuming that the decision boundary may lie in the low density region. Apart from considering these criterion alone, combinations of these criteria are also proposed in the literature.

Uncertainty

In the literature, several methods are presented to recognize the samples whose class assignment is most ambiguous by using the current classifier model. A family of methods uses posterior probability for determining uncertainty of a sample [121, 147, 190, 195]. Another family uses prediction results of a committee of classifiers [50, 53, 73, 86, 152, 166, 207, 229]. Another interesting family of methods is based on SVMs. They have been successful in many real-world learning tasks [26, 31, 45, 67, 147, 164, 200, 226, 253]. They are also categorized as large margin based AL methods [229].

The family of probability based methods analyze each sample of the unlabelled pool according to their estimated posterior probabilities of belonging to a class C_j (i.e., $p(C_j|x)$) and ranks the most uncertain sample. KL-max is a probability based method that tries to include the sample in training set whose inclusion maximizes the posterior probability distribution. In this direction, in [195] the Kullbach-Leibler divergence (KLD) between the distributions before and after adding the sample is maximized. In [190] same strategy is used for HSI classification exploiting maximum-likelihood classifier. To this end, test is conducted for each sample in unlabelled pool. Each sample is added to the training set one by one along with its class label and its posterior probability is maximized. The KLD is computed between the posterior distributions of the classifier model including and without including the sample. Once we have the KLD for all the samples in U, following function (Eq. 1.4) is computed and the sample with maximum value is considered most uncertain.

$$\hat{x}^{KLD-max} = \arg\max_{x_i \in U} \left\{ \sum_{C_k \in \{1,\ldots,c\}} \frac{1}{u-1} \times KLD(P^{+}(C_k|x)||P(C_k|x))P(y_i^* = C_k|x_i) \right\} \tag{1.4}$$

Where $u = |U|$ and KLD can be defined as shown in Eq. 1.5 considering the increased training set $L^+ = L \cup (x_i, y_i^*)$, where y_i^* is the class that maximizes the posterior probability.

$$KLD(P^+(C_k|x)||P(C_k|x)) = \sum_{x_j \in U \backslash x_i} P^+(y_j^* = C_k|x_j) \log \frac{P^+(y_j^* = C_k|x_j)}{P(y_j^* = C_k|x_j)} \tag{1.5}$$

Here, the $P^+(C_k|x)$ is the posterior distribution of class C_k estimated using L^+. In [121], the samples selected in previous iteration whose relevance has decreased now in current classifier model, were boosted. Such methods are computationally demanding and are practically applicable only with classifiers having small computational cost. Furthermore, These methods are designed to select one sample in each iteration and thus are not able to select a batch of samples. Another probability based method is breaking ties (BT) that selects a sample whose probability for belonging to all classes are nearly equal. In binary classification the decision is quite clear where the sample having near 0.5 probability for both the classes is considered as most uncertain. In other words, the sample having smallest difference between the two probabilities is selected as most uncertain sample. For multi-class scenario, this idea is extended to find the probability of belonging to each class and the difference of first two maximum probability values is considered. This can be formulated as following where C_k^+ is the class for which sample x_i has maximum probability.

$$\hat{x}^{BT} = \arg\min_{x_i \in U} \left\{ \max_{C_k \in N} \{p(y_i^* = C_k|x_i)\} - \max_{C_k \in N \backslash C_k^+} \{p(y_i^* = C_k|x_i)\} \right\} \tag{1.6}$$

The family of methods based on committee of classifiers predicts the class label for a sample considering multiple learning algorithms. The sample for which maximum disagreement exists in the predicted class assignments, is selected as most uncertain [53, 86, 207]. In [152], for constructing a committee, bagging [22] is suggested. To this end, q different classifier models are prepared using q different training sets randomly drawn from current labelled pool L. Class label for each sample in U is predicted using the q different classifiers to obtain q class labels for each sample. In order to determine the maximum disagreement computation of entropy is suggested in [228]. The entropy $H^{BAG}(x_i)$ for sample $x_i \in U$ is computed as follows.

$$H^{BAG}(x_i) = -\sum_{k=1}^{N_i} p^{BAG}(y_i^* = C_k|x_i) \log \left[p^{BAG}(y_i^* = C_k|x_i) \right] \tag{1.7}$$

where, $p^{BAG}(y_i^* = C_k|x_i)$ is the probability that the committee of q classifier models will predict C_k as the class level for sample x_i. This can be formulated as Eq. 1.8.

$$p^{BAG}(y_i^* = C_k|x_i) = \frac{\sum_{m=1}^{q} \delta(y_{i,m}^*, C_k)}{\sum_{m=1}^{q} \sum_{j=1}^{N_i} \delta(y_{i,m}^*, C_j)}. \tag{1.8}$$

In the Eq. 1.7 and 1.8, y_i^* is the predicted class label for pixel x_i, N_i is the number of classes predicted for sample $x_i, 1 \leq N_i \leq c$ and $\delta(y_{i,m}^*, C_j)$ is an operator that returns 1 if the mth classifier model predicts class label C_j for the sample x_i otherwise it returns 0. The value of entropy is maximum for maximum disagreement between classifier models and minimum when all the classifiers agree on a class label. In [50] the measure is bounded with respect to the number of classes predicted by the committee by normalizing it. The *normalized entropy query-by-bagging* (nEQB) can be formulated as follows.

$$\hat{x}^{nEQB} = \arg\max_{x_i \in U} \left\{ \frac{H^{BAG}(x_i)}{\log N_i} \right\} \tag{1.9}$$

Another committee based AL method is *adaptive maximum disagreement* (AMD). In this method, in place of considering q set of training samples, q subsets of feature space are considered [73, 166]. q different classifier models (each based on different subset of feature space) are trained to classify each of the samples in U. For each sample q predictions are obtained and the sample with maximum disagreement is selected as most uncertain. This can be computed similarly to the nEQB [229].

An interesting family of methods are based on SVMs [26, 31, 45, 67, 147, 164, 200, 226, 253]. SVMs creates a hyperplane to separate a class of samples from the rest of universe. SVM is by default a binary classifier which aims at dividing the d-dimensional feature space into two subspace, one subspace for each class. The decision hyperplane has an associated discrimination function $f(x) =< w.x > +b$, whose sign decides the class in which a sample x belongs to. The two possible classes in this case are +1 or -1. The SVMs uses kernel trick in which the samples are projected in higher dimensional feature space to better decide the decision hyperplane. These kernel functions $K(.)$ has to satisfy Mercer's conditions [30]. Training of the classifier model aims at calculating the values of Lagrange multipliers α_i related to the original training samples $x_i \in L$ using the Lagrange optimization theory. Once the training of the classifier model is complete, the following discrimination function is ready to predict an unlabelled samples x.

$$f(x) = \sum_{i=1}^{N_{SV}} y_i \alpha_i K(x_i \cdot x) + b \tag{1.10}$$

In Eq. 1.10, N_{SV} is the number of support vectors $(x_i \in L)$ used during the training of the classifier model, the y_i are the class labels of support vectors x_i associated to non zero α_i and the $K(x_i \cdot x)$ is the kernel function that determines the similarity between the sample x and the support vector x_i. In multi class scenario, an ensemble of binary SVM classifiers are combined according to some strategy [162]. The two strategies are one against all (OAA) and one against one (OAO). In the first strategy, for a c-class problem c binary classifiers are modeled one for each class and maximum positive distance from the hyperplane determines the class of a sample. In OAA the hyperplane separates samples of a class from rest of the universe. In the second strategy (i.e., OAO) classifier models are created for each pair of classes. The OAO is computationally demanding and thus OAA is preferred in research community. More detail on SVM in remote sensing can be found in [162]. Since the SVM classifier computes a distance for each sample to predict the corresponding class label, it is suitable to act as a base for active learning. The distance from the decision hyperplane can directly be used for deciding the classification confidence for an unlabelled sample. The samples present within the margin of the current classifier models are the probable support vectors of the future classifier model.

Margin Sampling (MS) is an AL method that attempts to select the unlabelled samples from U, which are nearest to the decision hyperplane [31, 200]. In [45, 253], the MS is designed in a way to minimize the selection of those samples which has less probability of becoming support vectors. In remote sensing MS was introduced in [164] which was latter improved in [26]. Basically, the unlabelled samples in U having lowest classification confidence (CC) are selected for labeling. The sample nearest to the hyperplane has lowest CC and thus it is selected as most uncertain. The MS can also be used in a multi-class scenario where the distances from c hyperplanes are recorded for c different classes and the minimum distance is considered as CC. This can be formulated as follows for a sample x and discrimination function f_i corresponding to ith class.

$$CC(x) = \min_{i=1,2\ldots c} \{|f_i(x)|\} \tag{1.11}$$

However in a multi-class scenario, consideration of the sample nearest to any one of the hyperplane may not lead to best result. The MS is further extended to a multi-class scenario with name multi-class level uncertainty (MCLU) [67]. The MCLU aims at identifying the sample that has maximum difference among the distances of first two farthest separating hyperplanes [232]. For this, the distance from each separating hyperplane is recorded, the two largest distances are noticed and the difference of these distances are considered as CC. For a sample x, MCLU can be formulated as follows.

$$
\begin{aligned}
r_{mx1} &= \arg \max_{i=1,2\ldots c} f_i(x) \\
r_{mx2} &= \arg \max_{\substack{j=1,2\ldots c \\ j \neq r_{mx1}}} f_j(x) \\
CC(x) &= f_{r_{mx1}}(x) - f_{r_{mx2}}(x)
\end{aligned}
\tag{1.12}
$$

Diversity

The main idea of diversity is to select the samples which are non-redundant. Addition of an informative sample to labelled pool will increase the training power while the redundant samples will add no value. Therefore, The samples which are diverse in nature along with having low classification confidence should be selected [67]. To this end, two diversity criteria are present in the literature: one is angle based diversity (ABD) and another is cluster based diversity (CBD).

The ABD suggests to measure the similarity of the samples based on the cosine angle distance. The cosine angle distance measures the distance between samples in the kernel space [23].

$$
Ang^{ABD}(x_i, x_j) = \frac{K(x_i, x_j)}{\sqrt{K(x_i, x_i)K(x_j, x_j)}} \tag{1.13}
$$

The angle $Ang^{ABD}(x_i, x_j)$ is small if x_i and x_j are close and it is a bigger value for large distance. The diversity algorithms have to select the samples which are at large distance, since the samples near to each other mostly represent redundant information.

The CBD groups the samples $x \in U$ into h clusters using standard k-means clustering algorithm [75] where h is the number of samples the algorithm has to select and finally, selects one representative sample from each of the cluster [239]. CBD focuses on the distribution of the data in feature space and since similar samples tend to group in one cluster, a representative from each cluster is enough for representing the whole distribution. An enhanced CBD (ECBD) is presented by Demir *et.al.* in [67] that uses kernel k-means [71, 98, 202, 208, 244] for clustering the samples to ensure better diversity.

Cluster assumption

The cluster assumption criteria aims at selecting the samples in the low density region of the feature space. The assumption is that the decision boundary usually lie in the low density region. The samples selected from this region are supposed to be more informative for class discrimination. In the literature, several works have attempted to select the samples from this reason [175, 176, 179]

Combination of criteria

In the literature, it is suggested to combine the above criteria to achieve better results [67, 176, 229]. The combination of uncertainty and diversity is quite popular in AL literature. This aims at selecting the samples which have low classification confidence and are non-redundant. For this, two different strategies are available in the literature. In the first strategy, while computing the classification confidence of each sample, weights β and $(1-\beta)$ are given to the two criteria. This is formulated in Eq. 1.14.

$$CC(x) = \beta(criterion1(x)) + (1-\beta)(criterion2(x)) \tag{1.14}$$

The second strategy is little different. In this strategy, m, $(m << |U|)$ samples are selected using $criterion1$ (usually uncertainty) and finally h $(h < m)$ samples are selected out of m selected samples based on $criterion2$ (usually diversity). The possible combinations suggested in the literature are MS-ABD, MS-CBD, MCLU-ABD, MCLU-CBD, MCLU-ECBD *etc.* Among these possible combinations, MCLU-ECBD is claimed to be performing better [67].

Spectral spatial AL technique

All the methods presented above considers the spectral values alone for determining informative pixels for hyperspectral classification. In the literature, few methods have recently attempted towards considering spatial information along with spectral values of the pixel while determining its information content [140, 174]. In [174] attempts are made towards selecting the samples which are spatially closest to current support vectors (with and without parzen window [76]), spatial entropy and combination of these criteria using multi-objective genetic algorithms with non-dominated sorting [213]. Li et al. [140] presents a framework, based on marginal probability distribution, that serves as an engine in which AL technique can exploit both spectral and spatial information. They have used posterior class probability, modelled with discriminative random field, in which the association potential is linked with a multinomial logistic regression classifier and the interaction potential is Markov random field multilevel logistic prior. The technique has a good accuracy but is computationally expensive and needs multiple parameter tuning using loopy belief propagation tools.

1.2.3 Integration of spectral-spatial information for analysis of HSI

Since the neighboring pixels have correlated information, integration of spectral and spatial information plays a significant role in classification of HSI [89, 91, 92, 191]. Spectral-spatial features are fusion of spectral and spatial information. For extracting spatial information, one well known method is the use of Markov random field (MRF) modelling. MRF is a family of probabilistic model and can be explained as a 2D stochastic process over discrete pixel lattices [70]. Several methods have considered MRF for extracting spatial information [4, 91, 165, 224]. Impediment of MRF being supervised in nature and having computation over-burden [83] lead us to unsupervised spatial information extraction. One of the unsupervised method for extraction of spatial information is segmentation. Classification accuracy may improve by segmentation of image into spatially homogeneous regions [89]. In the literature, a number of works have used segmentation for extraction of spatial information [92, 93, 157, 199]. Another unsupervised method for extraction of spatial information is based on mathematical morphology (MM). Ample number of works are available in literature which have used MM based spectral-spatial information [11, 13, 58, 63, 79, 187, 188, 196, 210, 227]. This thesis concentrates on spectral-spatial information integration based on MM framework.

In the MM framework, an original image is filtered by applying MM filtering techniques and the filtering results obtained by using a sequence of thresholds are concatenated with the original image to form a data structure called morphological profile. In MM, there are two basic operators namely, dilation and erosion. Other morphological filtering operations like opening and closing are composition of dilation and erosion.

MM operators and filters

Fundamental morphological operators are dilation and erosion [205, 206]. Dilation $\delta_E(I)$ in gray-scale image I, replaces the pixel intensity with the maximum intensity value present in its neighborhood which is defined by a structuring element (SE). SE is a small structure which sets boundary of neighborhood for a pixel to be investigated. Erosion $\varepsilon_E(I)$ is dual of dilation and replaces the pixels by minimum intensity.

Two important morphological filters are opening and closing. Opening of an image I, $\gamma_E(I)$, by a structuring element E is defined as the erosion of I followed by the dilation with the symmetrical structuring element E.

$$\gamma_E(I) = \delta_E[\varepsilon_E(I)] \tag{1.15}$$

Closing of an image I, $\phi_E(I)$, by a structuring element E is defined as the dilation of I by E followed by the erosion with symmetric SE.

$$\phi_E(I) = \varepsilon_E[\delta_E(I)] \tag{1.16}$$

When opening or closing is applied on image, structures smaller than SE disappears. These filters may introduce false structures or modify existing structures which can be avoided by geodesic reconstruction. The composition of erosion and reconstruction by dilation is called opening by reconstruction or geodesic opening. The composition of dilation and reconstruction by erosion is called closing by reconstruction or geodesic closing. On applying geodesic opening or closing we get a filtered image image with some preserved objects. Using varying size of SE we get multiple filtered images (called granulometry) preserving shapes and size of all objects present in the image. A granulometry generated by geodesic opening using SE of an increasing size is called opening profile (OP). Similarly, a granulometry generated by geodesic closing using SE of an increasing size is called closing profile (CP). OP of image I can be define as:

$$OP(I) = \{\gamma_R^{E_1}(I), \gamma_R^{E_2}(I), ..., \gamma_R^{E_t}(I)\} \tag{1.17}$$

where t is the number of opening by reconstructions. $\gamma_R^{E_i}(I)$ is opening by reconstruction considering the size of structuring element E_i. It is defined as:

$$\gamma_R^{E_i}(I) = R_\delta^{E_i}[\varepsilon_{E_i}(I)] \tag{1.18}$$

where $R_\delta^{E_i}$ is reconstruction by dilation. Similarly, CP of an image I can be defined as:

$$CP(I) = \{\phi_R^{E_1}(I), \phi_R^{E_2}(I), ..., \phi_R^{E_t}(I)\} \tag{1.19}$$

where $\phi_R^{E_i}(I)$ is closing by reconstruction

$$\phi_R^{E_i}(I) = R_\varepsilon^{E_i}[\delta_{E_i}(I)] \tag{1.20}$$

and $R_\varepsilon^{E_i}$ is reconstruction by erosion.

Morphological profile (MP)

MP of an image I is the concatenation of image I with its opening profile and its closing profile i.e.,

$$MP(I) = \{CP(I), I, OP(I)\}. \tag{1.21}$$

Thus, MP of image I is a collection of $2t + 1$ similar images with different spatial information [186]. For HSI one can integrate spectral-spatial features by generating MP for all bands and use them together, but it will increase the dimension exponentially. To mitigate this problem, one option is to reduce the dimension of the original HSI and then integrate the generated MP for each image in the reduced dimension. This is what we called extended morphological profile(EMP) [14].

For reducing the dimension any of the several unsupervised or supervised transformation techniques can be used. However,the most preferred technique in literature is principle component analysis (PCA). The dimensionality of HSI is reduced with help of PCA by selecting first l principle components (PCs). Then EMP for HSI H is generated by concatenating the MP of l different images generated for each PC.

$$EMP(H) = \{MP(PC_1), MP(PC_2), ..., MP(PC_l)\} \tag{1.22}$$

This results in $l(2t+1)$ images containing spectral and spatial information to represent the pixels of HSI. For each pixel in HSI we get a vector of size $l(2t+1)$ having spectral-spatial information. In [14], concept of EMP was successfully implemented for HSI using PCA for dimensionality reduction. Other supervised dimensionality reduction for generating EMP has also been considered in [82] and stacked EMP has been classified using an SVM. Different variants of MP are present in literature which considers difference of subsequent filtered images in each profile and generate derivative of MP (DMP) [114].

No doubt, MP is a powerful technique for extracting spatial features still some limitations can be found in MP which are as follows.

1. The shape of SE is fixed, which is being considered as main limitation in extracting information from a scene where objects for varying sizes are possible.
2. The SE based morphological filtering can not filter the image based on the properties of objects in the image.
3. Another limitation of MP is computation complexity. Intermediate or final result of one filtering operation is of no use in another operation. Hence, computational complexity increases linearly with number of filtering operations.

To address these issues, the concept of attribute profiles is proposed in [56] that is based on attribute filtering operation [21, 197].

Attribute filters (AFs)

AFs operate on connected components of image. The aim of this filter is to retain those connected components of the image which satisfy a threshold criterion [21]. For example, let attribute be area and A_r is the threshold value. For this configuration, the area filter will preserve those connected components of the image, which are having area greater than or equal to A_r. Other smaller components will merge with their local background. Several attributes like area, volume, gray-level homogeneity, shape descriptors, are suggested in the literature [197].

In greater details, Fig. 1-6(a) shows an example of area filtering by considering a simple gray-scale image. The original image can be seen as six flat regions or flat zones identified by {A0, B1, C2, D1, E2, F2}. The number followed by

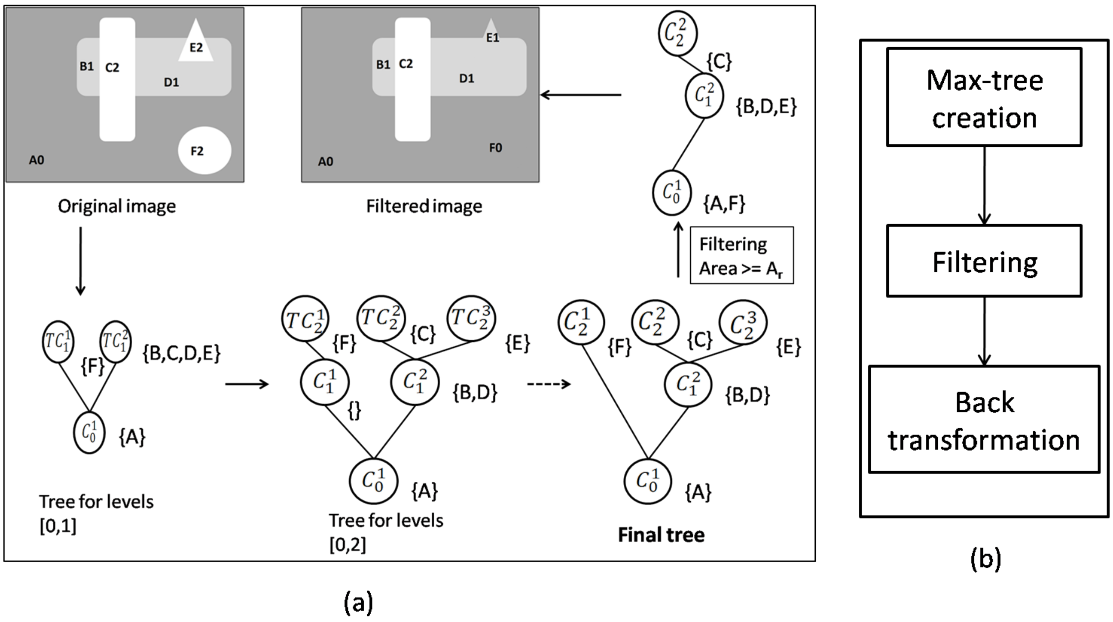

Figure 1-6: Max-tree construction and attribute filtering considering area as attribute.

the letter represents the gray-value of the region. An attribute filtering has three phases as shown in Fig. 1-6 (b). It starts with max-tree creation.

Max-tree creation: For this, a threshold h is set to 0 (minimum gray-value of the image) and the pixels of the image having gray-value 0 (region A) are assigned to the root node of tree i.e. C_0^1={A0}. The pixels having gray-values strictly higher than 0, formed two connected components that are temporarily assigned to two nodes: TC_1^1={F2} and TC_1^2={B1, C2, D1, E2}. This process of separating a subset of pixels from the pixels having strictly higher gray-values is called *binarization*. The first iteration creates the tree for gray-level [0,1]. In the next iteration, the threshold is increased by 1 i.e. $h = 1$. Now each temporary node is processed as separate gray-scale image. For this, binarization corresponding to each temporary node is performed and made them final retaining only the pixels having gray-value $h = 1$. For the remaining pixels, new temporary nodes are created which will be analyzed in next iteration. For instance, the temporary node TC_1^2 in Fig. 1-6(a) is binarized and C_1^2 is made final with pixels of region B1 and D1 which are having gray-value $h = 1$. The regions C2 and E2 formed two temporary nodes as TC_2^2={C2} and TC_2^3={E2}. The complete tree is constructed by iterating this process for all the temporary nodes at level h (h varies from the minimum to the maximum gray-values of the considered image). The tree constructed by this procedure may possess some empty nodes. Such nodes are removed at the end of the tree construction. The final constructed tree is called a *max-tree* which is the structural representation of a gray-scale image. For further details on max-tree creation readers may refer to [197].

Filtering: Once the construction of the max-tree is over, some specific criterion can be assessed on each node of the tree. Nodes satisfying the criteria are preserved and those do not are pruned. The pixels associated to the pruned node are assigned to its parent node. For example, in Fig. 1-6(a) the area criterion (Area $\geq A_r$) is assessed on each node and the node like C_2^1 and C_2^3 whose area is less than A_r are pruned. As a result, in filtered tree only three nodes are present namely C_0^1, C_1^2 and C_2^2. These nodes represent the connected components having area greater than the threshold value A_r.

Back transformation: At the end of the process, the filtered tree is transformed back to the gray-scale image by assigning gray-value h to the pixels of C_h^k, $\forall k$ and $\forall h$.

According to the above discussion filtering of a gray-scale image is a three steps process as illustrated in Fig. 1-6(b). First, a max-tree is constructed for the gray-scale image. Second, the constructed max-tree is filtered according to a specific threshold criterion based on a attribute. Finally, the filtered tree is transformed back to a gray-scale image with connected components satisfying the filtering criteria. In the example area has been considered as an attribute in filtering step. Other attributes like gray-level homogeneity, volume, size of bounding box, *etc.*, can also be used for filtering. In fact any measure that can be computed on the region of an image can act as an attribute [21]. Different filtering rules have been suggested in literature [197, 230]. Example illustrated in Fig. 1-6(a) is *area thinning* where the brighter objects having small area disappears from the image. Similarly to remove smaller dark areas, *area thickening* operation can be performed by filtering on min-tree or using duality property of morphological operators. Thinning and thickening operations can work with non-increasing attributes. If the attribute is increasing ($f \leq g \Leftrightarrow \psi(f) \leq \psi(g)$ for all gray-scale images f, g), the thinning (thickening) operation is called an opening (closing). Some of the increasing attributes are area, volume, size of bounding box *etc.*, and some of the non-increasing attributes are shape descriptors, gray-level homogeneity, orientation of region *etc.*

For applying attribute filtering operation a threshold is required during filtering process. Varying the threshold value one can have multiple filtering results for a gray-scale image. Thus, a thinning profile can be built for a gray-scale image by applying thinning filter with an increasing sequence of thresholds ($threshold_i < threshold_j \Leftrightarrow i < j$). Thinning profile ($TnP$) for a gray-scale image I with L

different thresholds can be defined as

$$TnP(I) = \{\gamma^{T_1}(I), \gamma^{T_2}(I), ..., \gamma^{T_L}(I)\} \tag{1.23}$$

where $\gamma^{T_i}(I)$ represents thinning of image I with i^{th} threshold. Similarly a thickening profile (TkP) can also be built for L different thresholds as

$$TkP(I) = \{\phi^{T_L}(I), \phi^{T_{L-1}}(I), ..., \phi^{T_2}(I), \phi^{T_1}(I)\} \tag{1.24}$$

where $\phi^{T_i}(I)$ represents thickening operation with i^{th} threshold.

Attribute profile (AP) and extended attribute profile (EAP)

AP of a gray-scale image is a concatenation of original image with thickening and thinning profiles computed with same set of thresholds.

$$AP(I) = \{TkP(I),\ I,\ TnP(I)\} \tag{1.25}$$

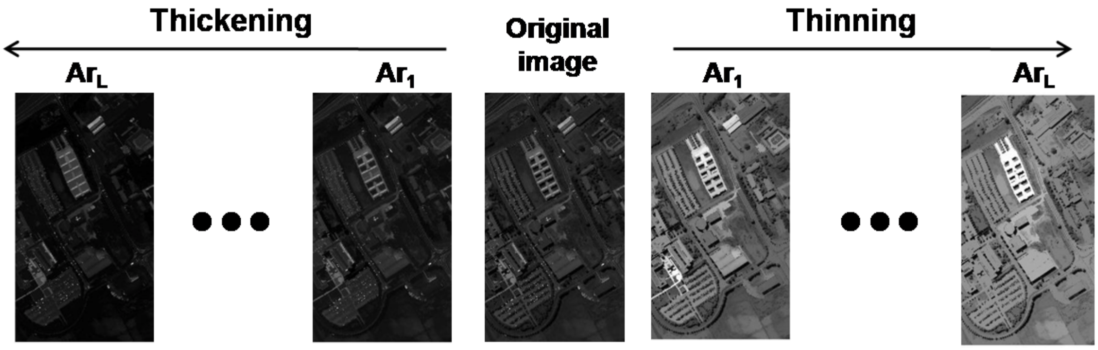

Figure 1-7: Attribute profile for a gray-scale image.

Fig. 1-7 shows an example of AP constructed for a gray-scale image. HSI is a stack of gray-scale images and we can generate spectral-spatial information based features by constructing AP for each of its gray-scale image. This will drastically increase the dimension of AP, resulting in increase of the computational burden and curse of dimensionality problem. To counter this issue, before finding AP the dimension of HSI is reduced mostly using principal component analysis (PCA) [120]. The first few principle components (PCs) that contain more than 99% information are selected. Then the APs constructed for each of the selected PC are concatenated to result in an EAP. An EAP for an HSI H can be represented as follows:

$$EAP(H) = \{AP(PC_1), AP(PC_2), ..., AP(PC_l)\} \tag{1.26}$$

where l is the number of selected PCs. Fig. 1-8 shows the construction of EAP for an HSI.

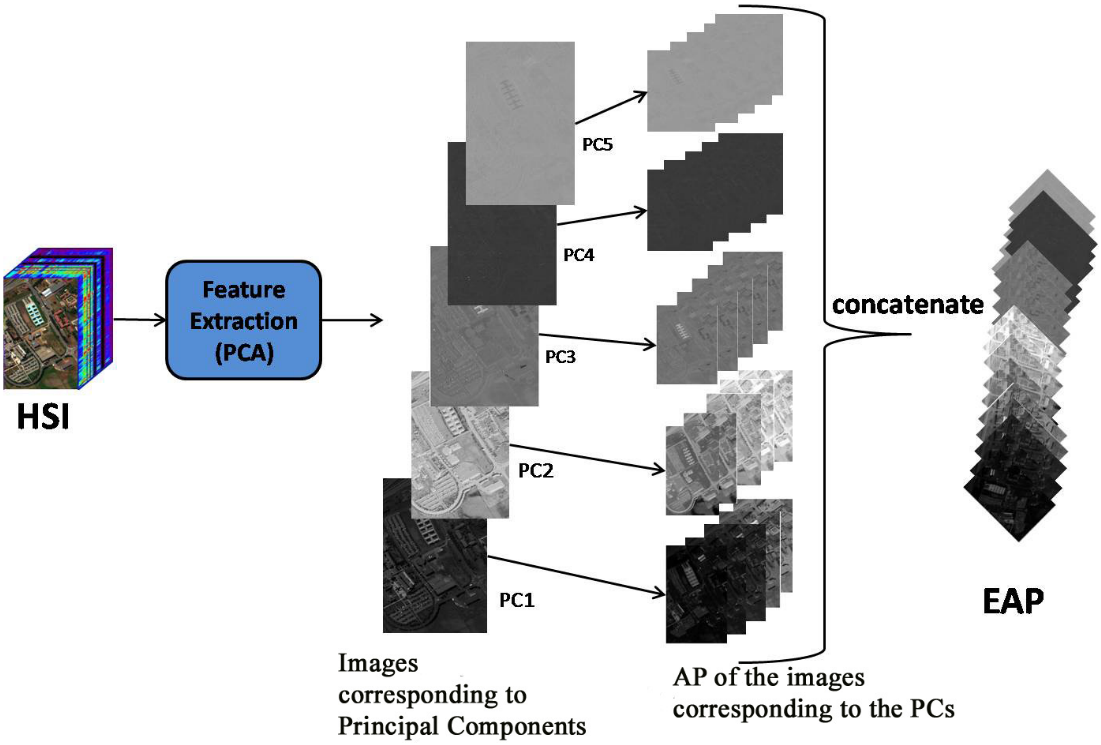
HSI
Feature Extraction (PCA)
PC5
PC4
PC3
PC2
PC1
Images corresponding to Principal Components
AP of the images corresponding to the PCs
concatenate
EAP

Figure 1-8: Extended attribute profile for an HSI.

Multi-attribute profiles

Generation of EAP considering a single attribute may have limitaed represention of proper spatial information. Multiple attributes that considered different spatial properties of the objects may be a better choice to incorporate proper spatial information. For an HSI H, an extended multi attribute profile (EMAP) can be constructed with r different attributes as follows:

$$EMAP(H) = \{EAP_{A1}(H), EAP'_{A2}(H), ..., EAP'_{Ar}(H)\} \tag{1.27}$$

Where $EAP_{A1}(H)$ is an EAP for the HSI H considering first attribute. It contains original PCs along with filtered images. $EAP'_{Ai}(H)$ represents EAP of H with i^{th} attribute and contains only filtered images. Fig. 1-9 represents the construction of EMAP for an HSI. Use of EAP and EMAP have shown significant improvement in classification accuracy. It also mitigated the limitation of SE based MPs since the shape of the connected components are adaptive and the max-tree is created only once. The use of EAP and EMAP for spectral-spatial classification is surveyed in [95].

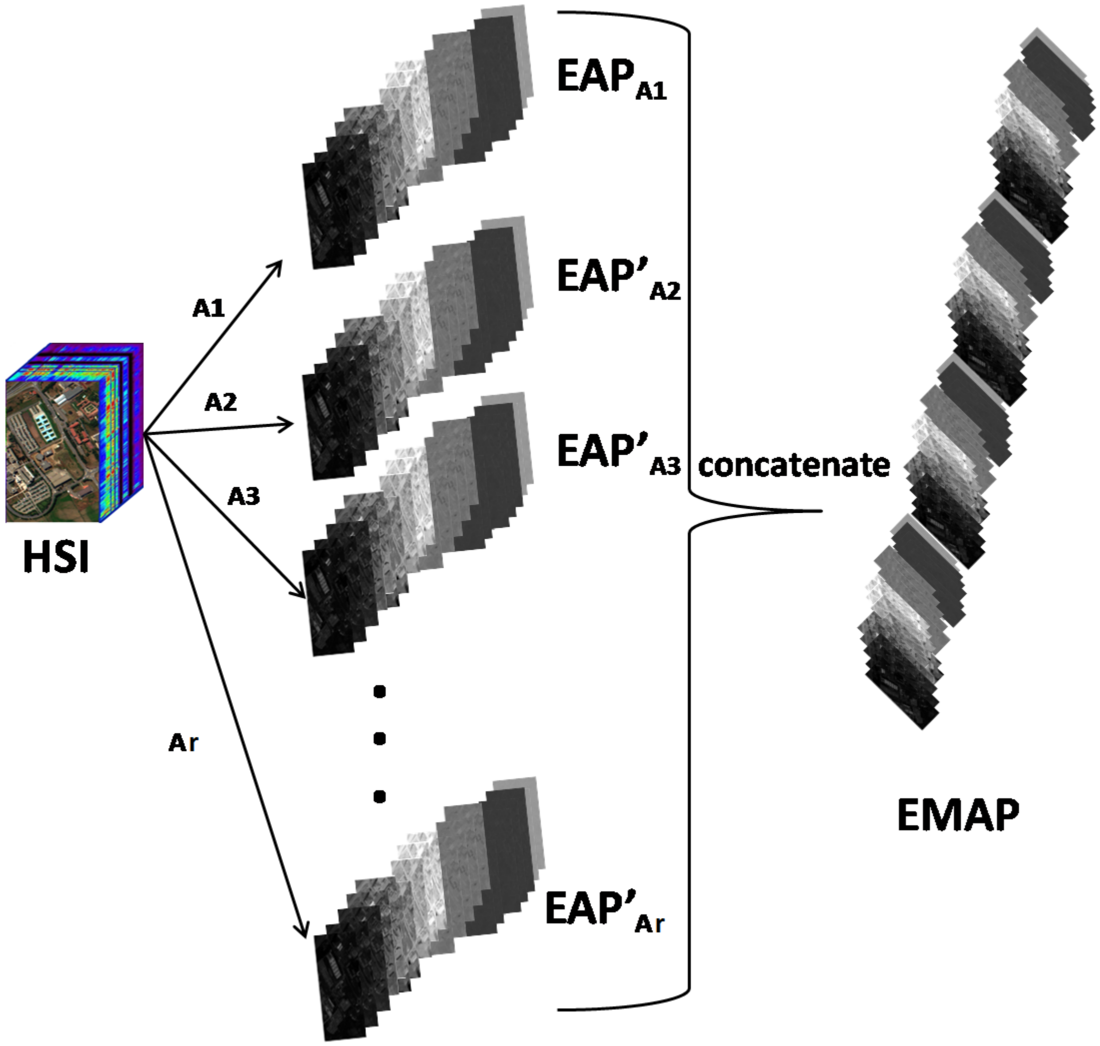

Figure 1-9: Extended multi attribute profile for an HSI.

When an EMAP is constructed using a dense sampling of filter parameter values from a wide range, it is called an entire extended multi attribute profile (EEMAP). Thus, the constructed EEMAP is rich in spaectral-spatial information but has large dimension and high redundancy.

1.3 Objective of the thesis

Mathematical morphology has already shown its potentiality in the integration of spectral and spatial information of HSI. Although, the classification results of HSI are significantly improved by integrating spectral and spatial information, at the same time it also introduces some critical challenges for effective analysis of HSI. The objective of the thesis is *to develop some robust techniques for effective analysis of HSIs using spectral-spatial information obtained by exploiting mathematical morphology.* Following are the three goals accomplished to achieve this objective.

1. *To develop a spectral-spatial multi-criteria active learning technique for classification of hyperspectral remote sensing images.* Training of a classifier model requires adequate number of labelled samples. On the other hand, in most of the HSI applications the number of available labeled samples are limited and also the generation of the labelled samples are costly. Active learning is a paradigm that reduces this labelling cost by selecting only informative samples for training the classifier. Although, a large number of AL methods are existing in the literature, most of them are proposed in spectral domain only. Few AL methods presented in spectral-spatial domain are computationally demanding. This goal of the research aims at developing a fast and robust spectral-spatial AL technique for classification of HSI.

2. *To develop an unsupervised technique for optimal feature selection in attribute profiles for spectral-spatial classification of hyperspectral images.* In mathematical morphology framework construction of multi-attribute profile is a promising approach to generate spectral-spatial features of HSI. In order to take into account the maximum spatial information, in the literature this multi-attribute profile is constructed considering a wide range of threshold values sampled in a small interval. As a result, the dimensionality of this constructed profile is huge and may introduce *curse of dimensionality* problem. In order to avoid such problem, a supervised feature selection method is proposed in the literature which is highly computational time demanding. There is no unsupervised feature selection method available in the literature for this purpose. This goal of the research aims at developing a fast unsupervised feature selection method to select the optimal subset of features from the constructed large multi-attribute profile for spectral-spatial classification of HSI.

3. *To develop a threshold-free attribute profile for spectral-spatial classification of hyperspectral images.* The attribute profile based spectral-spatial classification methods existing in the literature require a set of threshold values (selected either manually or automatically) for generating the filtered images. Since a single filtered image is unable to capture sufficient spatial information, multiple threshold values are used to generate several filtered images in an attribute profile. As a result, the construction of an attribute

profile is time consuming and may result in a high dimensionality. To the best of our knowledge, no method has been proposed in the HSI literature that generates attribute profiles without employing threshold values. This goal of the research aims at developing a fast method for constructing low dimensional attribute profiles that capture the maximum spatial information without using threshold parameter for spectral-spatial classification of HSI.

1.4 Organization of the thesis

The thesis is organized as follows:

- The chapter 1 introduces the thesis and the challenges of HSI classification highlighting dimensionality reduction, scarcity of labelled sample and integration of spectral and spatial information. It also presents the review on the state-of-the-art literature methods for mitigating such challenges. Furthermore, it lays down the objective of the thesis.
- Chapter 2 presents a novel multi-criteria AL technique that combines uncertainty, diversity and cluster assumption criteria by exploiting the properties of k-means clustering, K-nearest neighbors algorithm, support vector machines and genetic algorithms for designing the query function of AL. It incorporates spatial information using extended morphological profile [181]. Experiments on four benchmark hyperspectral data sets demonstrate that the proposed method outperforms five state-of-the-art active learning methods.
- Chapter 3 presents an unsupervised technique for optimal feature selection in attribute profiles for spectral-spatial classification of hyperspectral images [15]. In order to incorporate proper spatial information, the proposed technique constructs a large EMAP by varying the filter parameter values within a wide range. Then, to select an optimal subset of feature from EMAP, GAs are exploited by defining a novel objective function. The experiments conducted on four real hyperspectral data sets show the robustness of the proposed method in terms of computation time and classification accuracy.
- Chapter 4 presents a threshold-free attribute profile for spectral-spatial classification of hyperspectral images [16]. This chapter presents a novel filtering

approach that does not require the definition of threshold parameter to construct attribute profiles of HSI. To show the effectiveness of the proposed method, four real hyperspectral data sets are considered and the results are compared with many state-of-the-art spectral-spatial classification techniques.

- Chapter 5 concludes the thesis by summarizing the works done and also lists the possible further research in this area.