Contextual image classification with Support Vector Machine

Elham Goumehei
November, 2010

Contextual image classification with Support Vector Machine

by

Elham Goumehei

Thesis submitted to the Faculty of Geo-information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: Geoinformatics

Chair:	Prof. Dr. Ir. A. Stein
External examiner:	Dr.Ir. B.G.H. Gorte
Supervisor:	Dr. V.A. Tolpekin
Second supervisor:	Prof. Dr. Ir. A. Stein
Member:	J.P.G. Bakx MSc

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION OF THE UNIVERSITY OF TWENTE
ENSCHEDE, THE NETHERLANDS

Disclaimer

This document describes work undertaken as part of a programme of study at the Faculty of Geo-information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the University.

Abstract

Wide range of remotely sensed data obtained from different sensors is currently available. This data requires to be analyzed to get information. One way to analyze remote sensing data is classification. Choosing a suitable classification algorithm is important to efficiently use this large data set. Several approaches have been introduced where the contextual information is one of the applicable introduced models in classification of remote sensing data. For characterizing contextual information Markov Random Field (MRF) has been found an efficient tool. Application of MRF is based on maximum a posterior (MAP) estimation. It is employed as prior probability density function (p.d.f.). For the conditional p.d.f. often Maximum Likelihood Classification (MLC) is used where assumes classes are normally distributed. This assumption is not always a correct assumption. This research proposed a new MRF-SVM model that explores Support Vector Machine (SVM) instead of MLC. Since Implementation of SVM presented an improved classification results compare to other classifiers like MLC (Foody & Mathur, 2004; Foody, et al., 2006; Huang, et al., 2002; Pal, 2006; Pal & Mathur, 2005). The introduced model uses Simulated Annealing (SA) for energy minimization. Contribution of prior and conditional models was controlled by a smoothness parameter.

SVM offers some flexibility choice of penalty parameter value and a kernel function. Influence of these choices was considered. SVM assigns label to classes while in application of SVM as conditional p.d.f. class probabilities are required. To compute class probabilities for SVM Plott's theory was used. Using class probabilities from Plott's theory model was implemented on image synthesized from an agricultural scene. The accuracy of produced results was assessed by means of kappa coefficient (κ). In addition, reproducibility of results was evaluated by standard deviation of ten runs of model for ten different input images. Results indicate sufficient classification accuracy where the maximum κ is 0.95. During the procedure effect of class separability was investigated too. Also, performance of the model was compared to MRF based on normal distribution assumption. An illustration of MRF-SVM implementation on Synthetic Aperture Radar (SAR) image was presented to demonstrate applicability of the developed model for classification of real data.

In conclusion, the experimental results prove the effectiveness of the developed model. Performance of the model on synthetic data in terms of accuracy and reproducibility is acceptable. The model gives high κ value while use real images may reduce it. Also employed image is smooth that may positively affect classification accuracy. The strength of the model is observed through classification results of exponentially distributed classes. The results of MRF-SVM for both normally and exponentially distributed classes are nearly identical whereas MRF based on MLC does not behave similarly for data with different probability distributions. In terms of computational time, the new model has similar number of iteration as MRF based on MLC. The study shows that the MRF-SVM model is applicable for classification of remotely sensed data.

Key words: *Markov Random Field, Support Vector Machine, maximum a posterior, Maximum Likelihood, class probability, Simulated Annealing, class separability, class distribution, exponential distribution classes*.

Acknowledgements

All the praise and adoration to almighty God, the beneficent, the merciful as the everlasting leader and teacher in my entire life; without his divine grace, the prosperous achievement of this research was impossible.

My utmost grateful and appreciation goes to my supervisors, Dr. Valentyn Tolpekin and Prof. Dr. Alfred Stein for their precious guidance, critical advice and beneficial suggestions throughout the research period. All your encouragements helped me to focus on all parts of my research which I thank you.

My sincere thanks go to all lecturers of K.N.Toosi University of technology, the program director of JKIP, Dr. Behzad Vosooghi, GFM coordinator Mr. Gerrit C. Huurneman and to all administrative staff of ITC for their support and effort during my M.Sc. period.

I thank my classmates, for all these eighteen month interval togetherness which yield to durable friendship, fellow-feeling and unity.

My heartfelt appreciation goes to my honourable parents and lovely younger brothers for their warm affection, encouragement and support; without them all these would not have been achievable.

Table of contents

List of figures

List of tables

1. Introduction

1.1. Background and problem statement

Remote sensing is a valuable tool in many area of science which can help to study earth processes and solve environmental and socio-economic problems. Wide range of applicability is making it common in many fields. So these cause to remote sensing data become an important source and also using this source is considerable as well. Remote sensing provides information in the form of satellite images while the uses of remote sensing often require specific information on land cover. This information can be obtained by image classification. In fact, classification assigns label of a land cover class to pixels. There are two types of image classification: supervised and unsupervised classification.

Supervised classification requires user to define classes and select appropriate training samples. This approach is available in statistical methodology that is a quantitative analysis or non-statistical, geometric techniques which try to separate classes by surfaces (Richards & Jia, 2006). Statistical supervised classification uses an assumption about distribution for labeling classes. One of the most common statistical supervised classifiers is Maximum Likelihood based on the normal (Gaussian) distribution.

Maximum Likelihood classifier (MLC) based on Bayes formula calculates the probability of a pixel belonging to each class and assigns that pixel to the class with the highest probability. The problem of MLC is that uses assumption of normal distribution while some classes do not follow that distribution. For example radar images intensities are exponentially distributed or high resolution multispectral data like QuickBird image's DN values are non-normally distributed. Also multimodality of class distribution sometimes is a problem that causes MLC to fail. On the other hand, Gaussian distribution has a continuous space with an infinite range of data values while DN values in RS images are integer, distributed in a finite domain.

In recent years, more classifiers has been introduced like Support Vector Machine (SVM) which does not make assumptions about class distribution and are able to show a substantial improvements over traditional methods (Tso & Mather, 2009). Support Vector Machine classifier only uses those training samples that are on part of the edge of the class in feature space because it is based on fitting an optimal separating hyperplane between classes according to those training samples which are on border of classes. SVM generally uses less sample training while can get better accuracy in result (Foody & Mathur, 2004).

In recent years, there has been a trend for modeling prior probability of classification based on the concept of context (Tso & Mather, 2009). By using the context the aim is to generate a smooth image classification pattern. In other words, concept of contextual classification is that each pixel is treated in relation to its neighbors. One of the useful tools for characterizing contextual information is Markov Random field (MRF). Use of MRF produces a smooth classification result which is more suitable for many applications.

MRF is explored for normally distributed classes while SVM performs image classification well and is not dependent on class distribution assumption. But the main problem is that SVM is not a contextual classification. Integration of SVM and MRF might improve classification results. How these two methods can cooperate is the aim of this research.

1.2. Research objective

To develop contextual image classification method based on SVM and MRF.

1.2.1. Research questions

- How can SVM be integrated in MRF based contextual classification?
- How to compute class conditional probabilities with SVM?
- How to estimate MRF and SVM parameters for the MRF-SVM classification?
- How much this combination can improve accuracy of classification?
- What is the computational time of the developed method?
- Is the developed technique suitable for non-normally distributed class?

1.3. Methodology

The research would be start with a literature review on SVM and MRF contextual classification methods to know detail of characteristics, strength, and weakness of each technique in classification of remotely sensed data. Study will be intended to know how to compute class conditional probabilities from SVM technique.

Next step is defining posterior probability for implementation SVM-MRF method. Maximization of posterior probability can be performed by minimizing energy function. Major difficulty of energy minimization is that energy functions have many local minimum which increase computational cost (Boykov, et al., 2001). Graph cuts have two most popular algorithm called the swap move algorithm and the expansion move algorithm which can compute a strong local minimum (Szeliski, et al., 2006). Also estimation of parameters for both SVM and MRF methods is part of implementation new technique.

Then method will be tested for synthetic images which have normally and non-normally (e.g. Gamma, Exponential, Poisson, multimodal) distributed DN values. According to results, capability of method and its defects can be determined and technique can change in some parts.

To apply method on real images we need to select a study area that have to two conditions: 1) DN value of classes does not distribute normally, 2) require contextual classification, in other words, pixel based classification does not produce good results for them. Also images of this study area should be available.

Last step is validation of results that can be done according to error matrix, overall accuracy and Kappa coefficient.

1.4. Structure of the thesis

This thesis contains seven chapters. The *first chapter* describes the background, problem statement, research objectives of the research. The *second chapter* discusses some related works on MRF and SVM separately for classification of remotely sensed data. *Chapter three* provides detail information of the model used in this research. The theory of both MRF and SVM are described in this chapter. *Chapter four* provides information about data types used for this study and the adapted model. In *Chapter five* obtained results are presented. *Chapter six* discusses the results and provides comparison analyses of the applied models. And *Chapter seven* concludes and provides recommendation for future research.

2. Literature Review

This chapter reviews some works related to application of MRF and SVM in image analysis in the field of remote sensing. MRF and SVM are considered separately in section 2.1 and 2.2, respectively. Section 2.1 includes three parts that in each part application of MRF model by different researchers in the filed of image analysis in remote sensing is discussed and effect of this application is considered. In section 2.2 usages of SVM model in two parts is expressed that first part introduces study in the application of SVM model solely in remote sensing and second part reviews the combination of SVM and other methods that in recent years was pondered.

2.1. Contextual classification

Use of contextual information for image classification and segmentation became more popular in recent years (Tso & Mather, 2009). Markov Random Field (MRF) provides a convenient way to model context information. The practical use of MRF model is based on equivalence of MRF and Gibbs random field which was proved by Hammersley and Clifford. MRF theory often is used based on statistical methodology. Geman and Geman (Geman & Geman, 1984) proposed the idea to use maximum a priori (MAP) as statistical criterion and MRF together. Others developed the algorithm and many researchers applied it in different image analysis tasks (Barker & Rayner, 1997; Bruzzone & Prieto, 2000; Kasetkasem, et al., 2005; Solberg, et al., 1996; Tso & Mather, 1999).

2.1.1. Markov Random Field for image segmentation

Solving image segmentation problem by MRF model became popular, such as use it in other image analysis tasks. Supervised and unsupervised texture segmentation based on a hierarchical MRF model was proposed (Hu & Fahmy, 1991). The hierarchical MRF model uses the multi-level logistic model which is a particular Gibbs random field model for modeling region distribution, and the binomial model for modeling texture inside the regions. Then MAP problem stated as: the MRF model is the prior probability and an inhomogeneous random filed defines the conditional probability.

Barker and Rayner (1997) employed unsupervised segmentation algorithm based on MRF model for noisy images. The model applied MRF as prior probability and Pseudo-likelihood as conditional probability. Pseudo-Likelihood was implemented according to normal distribution where a Gaussian noise model was defined for each class. Results show improvement in accuracy.

MRF based on region adjacency graph (RAG) was implemented by (Sarkar, et al., 2000). The approach used an initially over segmented image as input image and defined MRF model over the RAG of the initially segmented regions. In a RAG, regions represent by nodes and arcs denote adjacency between regions. From the results it was expected that model can perform acceptable for any gray value image. So (Sarkar, et al., 2002) generated the defined model for multispectral images and results compared to Gaussian Maximum likelihood. There was improvement for obtained accuracy of proposed model compared to the maximum likelihood while no knowledge about image is

necessary for the proposed model. In 2007, Xia (Xia, et al., 2007) used the approach for synthetic aperture radar images and obtained an increased segmentation precision.

2.1.2. Markov Random Fields for remote sensing image classification

The wide range of data sets in remote sensing (RS) which differs in terms of spectral, spatial, and temporal resolution requires identification of suitable algorithms to use these RS data appropriately. Application of MRF as a contextual classification is an accepted approach in RS image classification.

Melgani and Serpico (Melgani & Serpico, 2003) used MRF model for a spatio-temporal classification. The approach applied a mutual MRF model which reduces the risk of propagating the classification error from one image taken at one time to image taken at another time. To define MAP problem they used separate contribution of three kinds of information: spectral information, spatial contextual and temporal contextual information. The model instead of normal class conditional probability density function used sensor-specific class conditional energy function by means of multilayer perceptron (MLP) neural networks. Model was implemented on Landsat TM and ERS-1 SAR images and acceptable accuracy was obtained.

Another study applied hidden MRF model for Radar images as unsupervised classification (Fjortoft, et al., 2003). This study used a generalized mixture estimation to determine the distribution families and parameters of classes. They investigated Gamma and K-distributed intensities. Results of the approach were remarkable but the problem was difficulty of the regularity parameter estimation.

(Tso & Olsen, 2005) proposed a MRF model with multi-scale fuzzy line process. Based on (Wei & Gertner, 2003) work which control contribution of edge pixels and obtained enhanced results, they used the wavelet-based edge detection method to extract multi-scale line features. Then they performed edge fusion on resulting multi-scale line feature to generate combined multi-scale fuzzy edge patterns for MRF model. In fact, in this model contextual effect will be turned off for detected edge pixel but for conditional probability density function they still used Gaussian distribution assumption. Results of this study shows improvement in accuracy of classification.

In 2005, Kasetkasem (Kasetkasem, et al., 2005) employed MRF model to generate Super-Resolution land (SRM) cover maps from remote sensing data. The approach assumed that super resolution map has MRF properties. This assumption removed a large number of misclassified pixels from obtained SRM. The efficiency of model was tested for Landsat ETM+ and IKONOS images and a significant increase in accuracies of produced SRMs was obtained while the model make normal distribution assumption for conditional density function.

Above mentioned studies are just some illustrations of several works done using MRF based on MAP criteria. There are many more researchers have performed MRF model for image analysis while the majority of these studies made normal distribution assumption for conditional probability density function which is not always a valid assumption for all data.

2.2. Support Vector Machines

Support vector machines (SVM) was developed by Vapnik in 1979 for image classification (Vapnik, 1995) and in 1995, he introduced the soft margin hyperplane for non-separable data which made SVM more applicable. Ability and good performance of SVM in variety of research domains make it attractive. SVM is gaining popularity in the field of remote sensing. SVM gives improved results with respect to traditional classifiers like maximum likelihood.

2.2.1. Apply SVM in hyperspectral remote sensing classification

In 1998, Gualtieri applied SVM for hyperspectral remote sensing classification (Gualtieri & Cromp, 1998). They used AVIRIS images and evaluated the performance of method for 4 and 16 classes. Due to ability of SVM for handling high dimensional data, study results were improved respect to traditional classifiers and represented potential to more works. Therefore, (Huang, et al., 2002) considered SVM to demonstrate the applicability of method for deriving land cover information from satellite images and compare it with the other classifiers like: maximum likelihood, neural network, and decision tree classifiers. The study investigated algorithm accuracy and stability for four methods. In terms of accuracy there were small differences in general, for algorithm stability SVM gave more stable overall accuracy with respect to other classifiers.

The effectiveness performance of SVM classification caused researchers to consider this classifier. Melgani and Bruzzone (Melgani & Bruzzone, 2002) applied SVM for AVIRIS hyperspectral data and compared SVM with K-nearest neighbors (K-nn) classifier and Radial Basis Functions (RBF) neural network. They used the original hyper-dimensional space to employ methods and got superior results for SVM approach. The assessment of results was based on the classification accuracy (overall accuracy), the computational time, and the stability of results. All the three assessment indicators prove the ability of SVM approach and encourage them to improve their experiment on SVM in the field of hyperspectral images (Melgani & Bruzzone, 2004).

Pal and Mathur (Pal & Mathur, 2005) also worked on SVM classification in remote sensing and considered two levels in their work; 1) they studied the effect of multi-class strategy on the performance of SVM and 2) compare the behavior of SVM with maximum likelihood (ML) and artificial neural network (ANN) on hyperspectral and multi-spectral data. First part suggested the use of “one against one” approach. For the second part, results reported higher accuracy for SVM classifier respect to ML and ANN even if the size of training dataset is small.

2.2.2. SVM model combined with other methods

When the efficiency of SVM in classification was proven through several studies, new approach was combined SVM with other methods. One of the first attempts was done by (Hermes, et al., 1999). The method was according to MAP criteria in Bayesian theorem for classifying Landsat TM images. They used a topological relation for prior probability and applied a method based on SVM for conditional probability density function. The produced results were considerably improved with small number of training data.

Substantial results of previous studies on SVM for classification of hyperspectral data as mentioned in section 2.2.1 encouraged Pal (Pal, 2006) to apply SVM with a Genetic Algorithm (GA) for feature

selection in land cover classification. Classification of hyperspectral data has two difficulties: 1) high correlation between bands, 2) estimation of more parameters. In this method GA utilized some of the existing bounds of the generalization error for SVM as the fitness functions. Their results for SVM/GA method were not desirable in computation cost.

Following the problem of feature selection in classification of hyperspectral data (Waske, et al., 2010) introduced another method using SVM. The proposed method is multiple classifier systems (Achlioptas, et al.) based on SVM and random feature selection (RFS), and results in terms of accuracy were compared with regular SVM. The MCS strategy used an RFS to perform various sets of feature subspaces, afterward, an individual SVM was applied on the feature subset to provide an individual classification result. These processes were performed based on number of classifiers and the classification output was combined with a majority vote. Experimental results gave significant improve in overall accuracy and more realistic results compared to standard SVM.

Application of combined SVM with other methods was not just in field of feature selection, the combination of SVM was considered in high spatial resolution classification too. A method was proposed that used optimized SVM as basis classifier and Random Forest (RF) to promote diversity which includes spatial information. The construction strategy was characterized by 1) a random mechanism to select subsets of training samples and spatial information in addition to spectral information, 2) SVM as basis classifier, and 3) the final map was generated by means of a weighed combination of classification maps. The proposed model achieved desirable accuracy and also visual quality of the classification (Waske, et al., 2010).

One more study was done by Bruzzone and Persello (Bruzzone & Persello, 2009) which presented a context-sensitive semi-supervised support vector machine (CS^4VM) model for classification problems in non-reliable training set. The proposed method aimed to exploit the information of the context patterns to reduce the bias effect of mislabeled patterns on the definition of separating hyperplanes of SVM. The strategy was based on supervised learning with original training sets and classification of context patterns using standard SVM with a neighborhood system. Then contextual semi-supervised learning was performed according to both original labeled patterns and semi-labeled context patterns with new method (CS^4VM). Results of standard SVM and new method classification for a very high resolution image (IKONOS images) and a medium resolution (Landsat) image were compared and robustness of new method was reported.

3. Methods

3.1. Markov Random Field classification based on Support Vector Machin

In image classification of remote sensing data estimating probability of classes is based on Bayesian theorem, which has had a strong and considerable influence on statistical modeling.

Bayes's theorem is one of the main tools for manipulating probabilities of any kind; where image classification uses this theorem as Bayesian classification. Bayes theory has two parts: prior and conditional probability density functions (p.d.f.). To estimate probability of classes, by use of combining these two functions (prior and conditional p.d.f.) maximum a posteriori (MAP) is expressed. In practice, availability of prior information is a problem for classification issue. Using Markov Random Field (MRF) as prior information becomes popular where it generates a smooth image classification pattern. MRF uses context information that may be derived from spectral, spatial, and even temporal attributes. Appropriate use of this context information can increase accuracy of classification (Magnussen, et al., 2004).

For conditional part of Bayesian theorem, the Maximum Likelihood (ML) has been widely adapted in remotely sensed image classification. ML assumes classes are normally distributed and then model the class-conditional p.d.f. while there are many image data (like QB and Radar images) in remote sensing with different distribution. This research uses Support Vector Machines (SVM) for conditional p.d.f. which makes no assumption about class distributions. In following theoretical explanation of both methods is described. Section 3.2 explains contextual classification and Markov Random Fields. And Support Vector Machine will be described in section 3.3 .

3.2. Contextual image classification

Contextual information is kind of spatial, temporal, and spectral relationship and is used for remotely sensed imagery interpretation in many studies. Contextual information, can be defined as how the probability of presence of one object (or objects) is affected by its (their) neighbors. In fact, in remote sensing classification, when a pixel is labeled as forest, it is likely to be surrounded by the same class of pixels unless that pixel is located in boundary area of that class. Markov Random Field (MRF) theory provides a convenient and consistent way to model this contextual information(Tso & Olsen, 2005). Such a modeling is the one which requires the least a priori information on the world model. Actually the simplest statistical model for an image consists of the probabilities of classes(Berthod, et al., 1996).

In following theoretical concepts of MRF will be expressed based on (Tso & Mather, 2009) and (Li, 2009).

3.2.1. Markov Random Field

Let $D = \{d_1, d_2, ..., d_m\}$ be a family of random variables on the set S with m sites in which each random variable D_i takes a value d_i in L. The family D is called random field. The set S is an image with m pixels; D is a set of pixel DN values. Also the label set L is a set of the user-defined information classes, e.g., L = {water, forest, pasture, or residential areas}. MRF is a model to describe dependencies of random variables.

According to defined random field, the configuration $w = \{w_1, w_2, ..., w_m\}$ for the set S is introduced, where $w_r \in L(1 \le r \le m)$. And the notation w is simplified to $w = \{w_1, w_2, ..., w_m\}$. Random field is satisfied the following three properties:

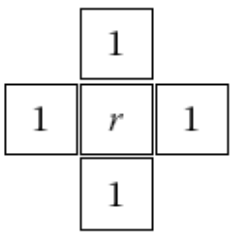
1) *Positivity*: $P(w) > 0$ for all possible configuration of w, $P(w)$ is the probability of configuration w. This can usually be satisfied in practice.

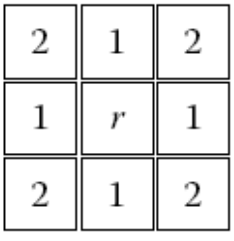
2) *Markovianity*: $P(w_r \mid w_{s-r}) = P(w_r \mid w_{Nr})$

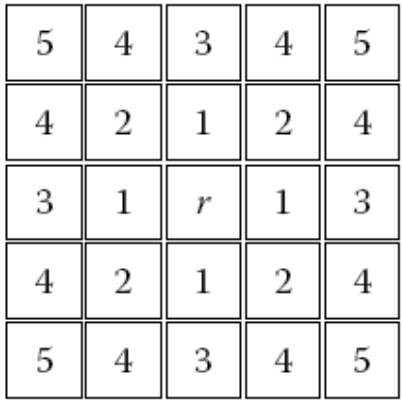
Where $S - r$ is the set difference, then w_{S-r} indicates the set labels at the sites in $S - r$ and Nr denotes the neighbors of site r. Markovianity indicates neighborhood interaction of sites. It means label of site r is directly dependent only on its neighbors.

3) *Homogeneity*: $P(w_r \mid w_{Nr})$ is the same for all sites r. This property states that probability for the label of site r does not depend on location of the site in S.

To define neighborhood in image analysis there is a system which specifies some surrounding pixels as neighbors. This usually system defines first-order neighbors with four pixels which have one common side with the given pixel, it is shown in Figure 3.1a. Second-order neighbors are pixels that have one corner in common with the pixel of interest (Figure 3.1b). Also higher-order neighbors can be defined that, up to five neighborhoods is shown in Figure 3.1c.







(a) (b) (c)

Figure 3.1 : neighborhood system to define neighbors for the pixel of interest, which (a) is the first-order neighbors with four pixels, (b) is the second-order neighbors that have a corner in common, and (c) is a higher-order one (up to five) in a similar manner. Source: (Tso & Mather, 2009)

3.2.2. Gibbs random field

Since MRF classifies pixels based on their neighbors, then describes the local properties of an image. While Gibbs Random Field (GRF) is defined as global model where gives the label to a specific pixel affected by the labels given to all other pixels. GRF characterizes an image in a global model by specifies a p.d.f. in the following form:

$$P(w) = \frac{1}{Z}\exp\left[-\frac{U(w)}{T}\right] \tag{3.1}$$

Where

$$Z = \sum_{\text{all possible configuration of } w} \exp\left[-\frac{U(w)}{T}\right] \tag{3.2}$$

Z is called the partition function and is the sum of all possible configuration of w. $U(w)$ is called energy function, and T is a constant named temperature. Based on Equation 3.1 maximization of $P(w)$ is equivalent to minimization of $U(w)$. The energy function is:

$$U(w) = \sum_{c \in C} V_c(w) \tag{3.3}$$

Where $V_c(w)$ is called the potential function with respect to clique type C. Clique C is a subset of image that indicates mutual neighborhood of all pairs of sites. Clique type of first order is shown in Figure 3.2.

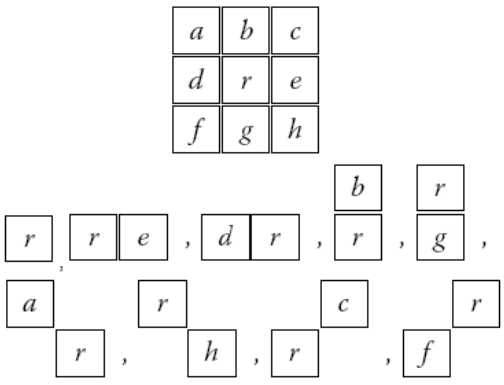


Figure 3.2 : all possible cliques with the neighborhood system on the first order for pixel of interest r .

According to definition of clique energy function expressed as:

$$U(w) = \sum_{\{r\} \in C_1} V_1(w_r) + \sum_{\{r, r'\} \in C_2} V_2(w_r, w_{r'}) + \ldots \tag{3.4}$$

By increasing the order of neighborhood system, the number of cliques expanded, and complexity grows too.

3.2.3. MRF-GRF equivalence

As mentioned above MRF describes local properties of an image and GRF is a global model. Based on Hammersley-Clifford theorem which describes equivalence of GRF and MRF for every MRF there is a unique GRF since GRF is defined in terms of cliques on neighborhood system. The proof of this theorem is presented in many researches for example in (Tso & Mather, 2009). This equivalence provides a simple way to dealing specify the MRF model by means of GRF model formulation.

3.2.4. Energy minimization

According to Equation 3.1, to get a maximum for $P(w)$, the energy function has to be minimized. To find an optimal solution for this minimization problem some iterative search techniques is used which seeks for minima. But always there is local and global minimum in solution space where convexity analysis may be used to solve the problem. In convexity analysis, if the energy function is convex (bowl-shaped with one minimum), based on Bayesian formulation; maximum a posteriori (MAP) for MRF can be obtained using basic search approach because of one minimum point in solution space. For non-convex energy functions which may have many local minima, finding a global minimum needs to search all local minimas. Tso and Mathur (Tso & Mather, 2009) have mentioned three algorithms with an iterative process: Simulated Annealing (SA), Iterated Conditional Modes (ICM), and Maximizer of Posterior Marginals (MPM). For this research SA is used which is explained in following.

3.2.4.1. Simulated Annealing (SA)

Simulated annealing is a stochastic method to find a global minimum solution. This technique increases the temperature parameter from a higher value to a low value during the iterative minimization. At a high temperature, it can locate the unique minimum where energy landscape is convex and smooth. By tracking the minimum, the energy is gradually decreased to get a sufficient low value. In fact, SA simulates a physical annealing procedure in which a metal structure is melted and then slowly is cooled down to make sure it has enough time to be hardened.

For intuition that how samples of w distributed in D, consider a system where any w in configuration space D has following probability:

$$P_T(w) = [P(w)]^{1/T} \tag{3.5}$$

Where $T > 0$ is the temperature parameter. $P_T(w)$ is concentrated on the peaks of $P(w)$, when $T \to 0$; for $T \to \infty$, $P_T(w)$ is a uniform distribution on D; and as $T = 1$, $P_T(w) = P(w)$.

```
Initialize T and w
repeat
        randomly sample w from P(w) under T ;
        decrease T ;
until (T → 0);
return w;
```

Figure 3.3: the simulated annealing algorithm (SA). source:(Li, 2009)

Figure 3.3 shows description of SA algorithm where $\mathrm{N}(w)$ is the neighborhood system. The algorithm first set a high value for T and w is set for a random configuration. At a fixed T, SA applies a sampling algorithm, such as Metropolis algorithm (Metropolis, et al., 1953) or Gibbs sampler (Geman & Geman, 1984) to converge to the equilibrium at the current temperature T. Then according to a carefully defined cooling schedule, T is decreased. This process is continued until the system becomes frozen (T is close to 0) in which energy function is near the minimum. Tuning the annealing schedule (cooling schedule) is very important where it has a critical effect on success of SA. Choosing an optimal schedule depends on the type and also on the size of the problem (Li, et al., 1997).

For two mentioned convergence theorems (Geman & Geman, 1984) presents proofs. The first considers the convergence of Metropolis's algorithm where state, the distribution of generated configuration is guaranteed to converge to the Gibbs distribution, if each configuration w is visited infinitely often. The second theorem is about temperature T of SA. It states if steps of decreasing temperature satisfies:

$$\lim_{k \to \infty} T_k = 0 \tag{3.6}$$

and

$$T_k \geq \frac{m \times \Delta}{\ln(1+k)} \tag{3.7}$$

Where $\Delta = \max_w U(w) - \min_w U(w)$, then the convergence can be guaranteed. This equation is very slow for practical applications, therefore faster cooling functioned may apply which is mentioned in (Tso & Mather, 2009) and (Li, 2009).

3.3. Support Vector Machine

Support vector machine was introduced in 1992 to generate maximal margin for non separable training data in feature space by hyperplanes (Vapnik, 2006). SVM was a binary classifier primitively which labeled classes as $+1$ and -1. The idea for this classification is separating these two classes with maximum margin. SVM constructs an optimal hyperplane for getting to maximized margin (Tso & Mather, 2009). In fact hyperplanes are decision boundaries for separating classes in feature space

which use training samples that lie on the edge of class distribution. Figure 3.4 shows how optimal hyperplane divide two classes based on maximum margin. Most materials in this section follow (Tso & Mather, 2009), (Cortes & Vapnik, 1995) and (Vapnik, 2006).

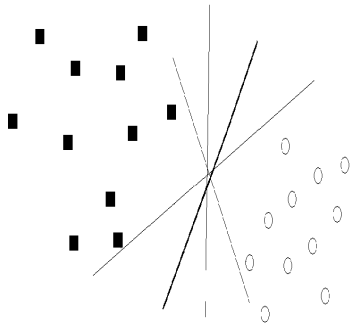


Figure 3.4 : several hyperplanes maybe used to separate samples.

3.3.1. Linear separable classification

Suppose the training data set is represented by pairs $\{x_i, y_i\}$, $i = 1,...,n$, $y_i \in \{1,-1\}$, $x_i \in R^d$ is a d-dimensional space, y_i is the label of class for training sample i that represent class $+1$ or -1. A hyperplane in feature space is defined by following equation:

$$w \cdot x + b = 0 \tag{3.8}$$

Where x is a point lying on the hyperplane, w is normal to the hyperplane and b indicates bias Figure 3.5. Separating hyperplane is defined for two classes like this:

$$w \cdot x_i + b \geq +1 \quad \text{for class } y_i = +1 \tag{3.9}$$

$$w \cdot x_i + b \leq -1 \quad \text{for class } y_i = -1 \tag{3.10}$$

All training data are supposed to satisfy two Equations 3.9 and 3.10 that are illustrated in Figure 3.5. These two equations can be combined to give:

$$y_i(w \cdot x_i + b) - 1 \geq 0 \tag{3.11}$$

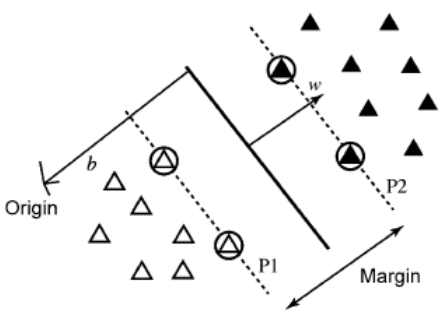


Figure 3.5 : SVM use support vectors to construct optimal hyperplane. Circled cases are support vectors. Source: (Foody & Mathur, 2004)

Training samples lie on two hyperplanes P1 and P2 are called support vectors and are parallel to the optimal hyperplane. Margin between these two is $\frac{2}{\|\mathrm{w}\|}$ and analysis aims to maximize this margin. The maximization of this margin leads to:

$$\min\left\{\frac{\| \omega \|^2}{2}\right\} \tag{3.12}$$

subject to the inequality constraints in Equation 3.11. Lagrange formulation of the Equation 3.11 is used which makes it easier to handle because of introducing, α_i, positive Lagrange multipliers (Burges, 1998). The primal Lagrange formulation gives:

$$L_p = \frac{1}{2} \| \mathrm{w} \|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\mathrm{w} \cdot \mathrm{x}_i + b \right) - 1 \right) + \sum_{i=1}^{n} \alpha_i \tag{3.13}$$

Which should be minimized with respect to w and b. To achieve that, derivative of L_p should be equal to zero:

$$\frac{\partial L_p}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.14}$$

$$\frac{\partial L_p}{\partial \mathrm{w}} = 0 \rightarrow \mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i \tag{3.15}$$

Substitution of 3.14 and 3.15 into equation 3.13 gives a dual Lagrangian L_D:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \mathrm{x}_i \cdot \mathrm{x}_j \tag{3.16}$$

Then the decision rule that separate training samples can be written as:

$$f\left(\mathrm{x}\right) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \left(\mathrm{x} \cdot \mathrm{x} \right) + b \right) \tag{3.17}$$

Now L_D have to be maximized respect to $\alpha_i \geq 0$, subject to constraint in Equation 3.14 with solution given by equation 3.15. Consider that there is Lagrange multiplier α_i for every training point. Just

support vector points have nonzero α_i, which are either exactly at the class boundaries or on the wrong side of the class boundaries (margin errors).

Notice that Lagrange formulas have different labels (p for primal, D for dual) to emphasize on their difference. Both of them have the same objective function but with different constraints; and the solution is found by minimizing L_p or by maximizing L_D.

This formulation of the SVM optimization problem is called the hard margin formulation; since training samples are classified without any error (classes are fully separable). While information classes of remotely sensed data are not usually separable.

3.3.2. Linear non-separable classification

Consider the case where classes are not fully separable. In this case one may want to have classes with some errors. In fact, training samples can be classified with a minimal number of errors. This sometimes is called soft margin method (Cortes & Vapnik, 1995; Tso & Mather, 2009). For this type of non separable data two types of SVM is introduced: C-SVM and nu-SVM which will be explained in the following.

C-SVM

To treat erroneous training samples (Cortes & Vapnik, 1995) defined non-negative variables $\xi_i \geq 0,\ i = 1,...,n$ are called slack variables, indicating distance of the sample from hyperplane of class it belongs to Figure 3.6. In fact ξ is introduced to classify non separable classes with a number of errors. Then the Equation 3.4 can be written as:

$$y_i\left(\mathrm{w} \cdot \mathrm{x}_i + b\right) \geq 1 - \xi_i \tag{3.18}$$

For outliers that exist in data set, a penalty term, $\mathrm{C}\sum_{i=1}^{n}\xi_i$ is added to penalize solution for very large ξ. The parameter C controls magnitude the penalty of training samples that lie on the wrong side of the hyperplane. Larger value of C leads to overfitting and decreases generalization capability (Tso & Mather, 2009). This type of SVM is usually called C_SVM.

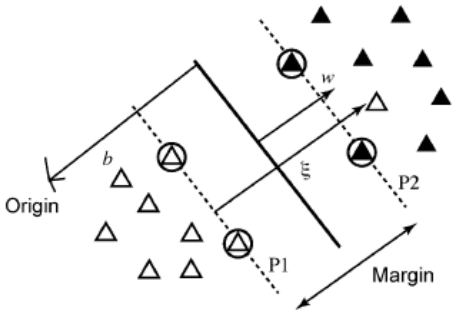


Figure 3.6: representing non-separable cases (Foody & Mathur, 2004)

Now, the optimal problem resolves to:

$$\min\left\{\frac{\| \mathrm{w} \|^2}{2} + C\sum_{i=1}^{n}\xi_i\right\} \tag{3.19}$$

Subject to the inequality constraint in Equation 3.18. The first term of Equation 3.19 is for maximizing the margin while the second part seeks to penalize training samples on the wrong sides of class boundaries in the non separable cases.

The above minimization is a quadratic objective function with linear constraints that is a standard problem in optimization theory (Belousov, et al., 2002). It can be solved by applying Lagrange theory. Then the primal Lagrange formulation is:

$$L_p = \frac{1}{2}\| \mathrm{w} \|^2 - \sum_{i=1}^{n}\alpha_i\left(y_i\left(\mathrm{w}\cdot \mathrm{x}_i + b\right) - 1 + \xi_i\right) + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\mu_i\xi_i \tag{3.20}$$

Where, μ_i are the Lagrange multipliers to enforce positivity of slack variables ξ_i. Now, Equation 3.20 should be minimized with respect to w, ξ and b. Using the solutions for substitution in Equation 3.20, maximize new equation:

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathrm{x}_i \cdot \mathrm{x}_j \tag{3.21}$$

Subject to

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C \text{ for } i = 1,\ldots,n$$

Now C is an upper bound for Lagrange multiplier α_i to enforce that any given support vector is allowed to exert on hyperplanes P1 or P2 .

ν-SVM

(Schölkopf, et al., 2000) introduced a new algorithm of SVM which defined parameter ν. Here they substituted C by new parameter ν:

$$\min\left\{\frac{1}{2}\| \mathrm{w} \|^2 - \nu\rho + \sum_{i=1}^{n}\xi_i\right\} \tag{3.22}$$

subject to

$$y_i(\mathrm{w} \cdot \mathrm{x}_i + b) \geq \rho - \xi_i \tag{3.23}$$

Where,

$$\xi_i \geq 0,\ \rho \geq 0$$

Following the same derivation as in Subsection 0 gives:

$$L_D = -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathrm{x}_i \cdot \mathrm{x}_j \tag{3.24}$$

Subject to

$$0 \leq \alpha_i \leq \frac{1}{n} \tag{3.25}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.26}$$

$$\sum_{i=1}^{n} \alpha_i \geq \nu \tag{3.27}$$

This Equation 3.24 compared to Equation 3.21 has two differences. First there is an additional constraint ν. Second, the linear term $\sum_{i=1}^{n} \alpha_i$ no appears in the objective function which is now quadratically homogeneous in α (Schölkopf, et al., 2000).

The new parameter ν controls support vectors and errors both as a single parameter. Controlling the number of support vectors has consequences for: 1) complexity of run-time, since evaluation time has a linear relationship with number of support vectors (Burges, 1998); 2) training time; 3) parameter ν is enough to train the algorithm only on the support vectors; 4) generalization error bounds.
Note that two algorithms are not different fundamentally while (Schölkopf, et al., 2000) showed that for certain parameter setting, results are similar.

3.3.3. Non-linear classification

To generalize the above method for non-linear cases, support vector machine uses an implicit mapping function Φ to map the input vector $\mathrm{x} \in R^n$ into a high-dimensional feature space H and constructs the optimal separating hyperplane in that space Figure 3.7:

$$\Phi : R^n \rightarrow H \tag{3.28}$$

A vector x in the feature space is represented as $\Phi(\mathrm{x})$ in the high-dimensional space H. Since problem optimization of SVM only uses dot products of two vectors, the method applies kernel function such that:

$$K(\mathrm{x}_i, \mathrm{x}_j) = (\Phi(\mathrm{x}_i), \Phi(\mathrm{x}_j)) \tag{3.29}$$

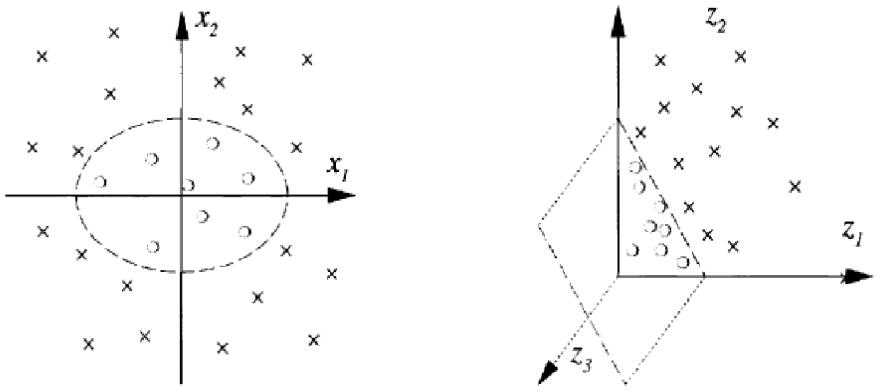


Figure 3.7: kernel functions map training samples into a higher dimensional space to find an appropriate decision boundary. Source:(Schölkopf & Smola, 2001)

Now, we can train SVM and use these kernels in a high-dimensional space. And the optimization problem can be written as:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathrm{x}_i, \mathrm{x}_j) \tag{3.30}$$

And decision rule now generalized into:

$$f(\mathrm{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathrm{x}, \mathrm{x}_i) + b \right) \tag{3.31}$$

The interesting fact of using kernel functions is that we do not need to know the explicit form of Φ (Cristianini & Shawe-Taylor, 2000). The kernel used must satisfy Mercer's condition. Then any (conditionally) positive definite function $K(\mathrm{x}, \mathrm{x}_i)$ can be used to construct a support vector machine(Vapnik, 2006).

3.3.3.1. Kernel functions

Kernels make it possible to map the data into a high-dimensional space to separate training samples by an appropriate decision boundary. The key is finding appropriate function that can be evaluate efficiently. Common kernels that are used for SVM method include the following:

- Linear kernel

$$K(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{x}_i^{\mathrm{T}} \mathrm{x}_j$$

- The Gaussian Radial Basis Function (RBF) kernel

$$K(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(- \frac{\| \mathrm{x}_i - \mathrm{x}_j \|^2}{2\sigma^2} \right)$$

Each of these kernel functions is used in different condition for example (Karatzoglou, et al., 2006) mention that RBF kernel is used when there is no prior knowledge about data.

Moreover, (Tso & Mather, 2009) discussed in their book impact of kernel function on SVM results and declared there is a close relation between choosing kernel function and performance of SVM. Also (Huang, et al., 2002) investigated impact of kernel function on the performance of SVM and results revealed that kernel type and kernel parameter can influence on shape of decision boundaries and subsequently on SVM performance.

4. Implementation

This chapter describes applied data and the processing steps to implement the MRF-SVM model are given. Section 4.1 describes the data sets that are used. Section 4.2 explains software and packages used to implement model. Methods, their characteristics, and parameters that were performed are explained in section 4.3.

4.1. Synthetic image

Synthetic image is an artificial form of real data which provides a useful source to improve understanding the complexity of reality. It also can be employed to emphasize the desirable aspects. In this case, controlling the probability density function of classes and their separability are desired. These reasons lead the research to apply synthetic images for implementing new method and study different aspects of introduced model.

In this study, an image with 60×60 pixels is synthesized. The image has two classes and single band. For each test, one image is generated which means every experiment uses new synthetic image. During image generation class separability was controlled. In the following, reference data and employed images will be explained.

4.1.1. Reference image

Reference map is a subset of real image. The image contains two homogenous classes. Figure 4.1 shows reference data.

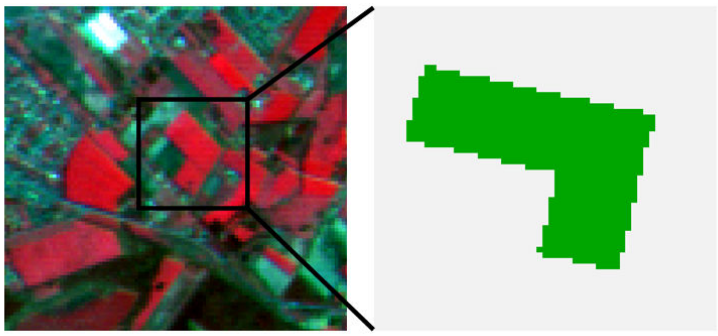
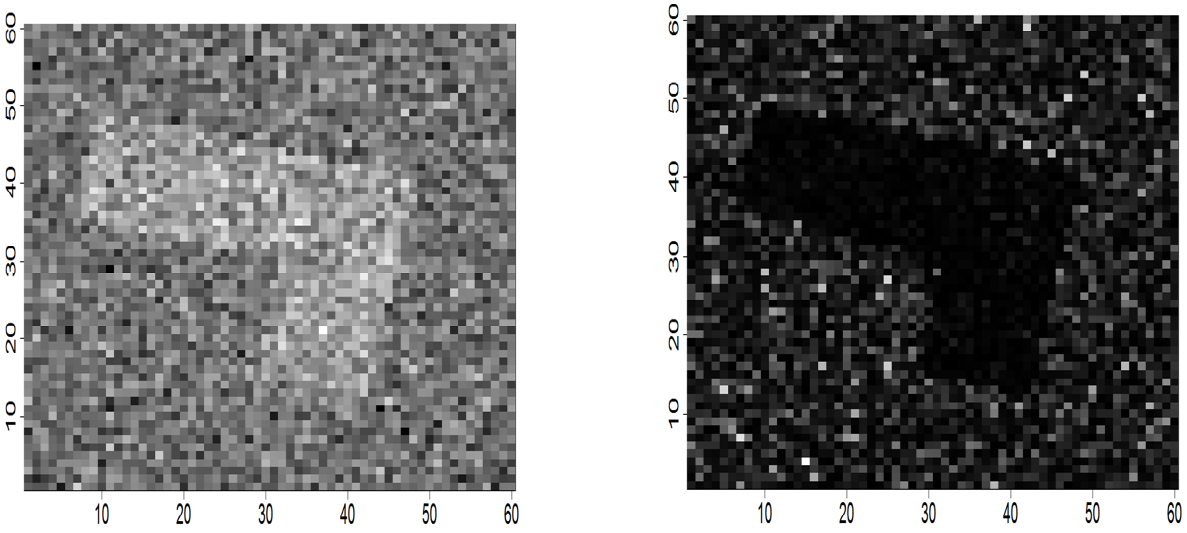


Figure 4.1: Left picture shows Landsat image that reference data was generated based on that and in right, reference image is presented

4.1.2. Generation of pixel values

Based on reference image pixel values were produced through random number generator using class parameters. Two types of image were generated. One with normally distributed classes and the other one with exponentially distributed classes. Throughout the thesis to simplify, image with normal distribution classes will be called ND image and image with exponentially distributed classes will be called ED image. An example of each class is shown in Figure 4.2a and b.



(a) (b)

Figure 4.2: (a) Image with normal distribution classes (ND image). (b) Image with exponential distribution classes (ED image).

Class separability for images was considered too which will be in the following. For implementing the model JM=0.5 was used that shows big overlap for class distributions.

4.2. Software

The introduced model was performed in R software, version 2.11.1. It is a programming language for statistical computing. R provides a wide variety of statistical techniques.

R introduced four packages to implement SVM. The first package in R for SVM was *e1071*. *Kernlab*, *klaR* and *svmpath* are the other packages in R which implement SVM. (Karatzoglou, et al., 2006) discussed these packages and compared their performance for support vector machine. This research uses kernlab package to implement SVM.

Kernlab is an extensive and flexible package for kernel-based learning methods in R(Karatzoglou, et al., 2004). *Kernlab* package uses ksvm() function to implement SVM that includes most important formulations and kernels for SVM and even let the user to define kernels. This function includes the C-SVM and ν-SVM classification algorithm to implement SVM. The *kernlab* package supports different kernel functions include linear kernel function and the radial basis function (RBF) which is used in this research. The package also has the ability to estimate automatically σ for RBF kernel function. In addition, *kernlab* package can produce class probability output instead of class labels

which is the main reason to select this package. In fact, to apply SVM model as conditional probability it is needed to use class probabilities. This package uses Platt's a posteriori probabilities (Karatzoglou, et al., 2006) to compute class probability that will be explained in section 4.3.2.

In addition, MRF codes were available in R and the SVM code was developed in this research.

4.3. Methods

As mentioned, methods in this research are based on Bayes formula which has two parts: prior and conditional probability density function (p.d.f.). To classify the input data MAP solution is adapted for the given dataset d and the class label w (chapter 3):

$$P(w_i \mid d_i) \propto P(w_i \mid w_{Ni})P(d_i \mid w_i) \tag{4.1}$$

Due to MRF and Gibbs random field equivalence Equation 4.1 can be written as:

$$U(w_i \mid d_i) = U(w_i \mid w_{Ni}) + U(d_i \mid w_i) \tag{4.2}$$

Where, $U(w_i \mid w_{Ni})$ is the prior energy function for neighborhood system Ni, $U(d_i \mid w_i)$ denotes the conditional energy and $U(w_i \mid d_i)$ is the posterior energy for one pixel. An additional parameter λ is defined to control contribution of prior and conditional energy function:

$$U(w_i \mid d_i) = \lambda \cdot U(w_i \mid w_{Ni}) + (1-\lambda) \cdot U(d_i \mid w_i) \tag{4.3}$$

The value of parameter λ is between 0 and 1. If $\lambda = 0$ the prior model (here the prior model is MRF) is completely ignored and if it becomes equal to one just the prior model is considered. Since this research is going to study the combined models (MRF based on SVM or ML), the value of smoothness parameter λ should not be equal to 0 or 1.

The posterior energy for the entire image is defined as:

$$U(w \mid d) = \sum_{i=1}^{n} U(w_i \mid d_i) \tag{4.4}$$

For the prior model MRF method is adapted and for the conditional p.d.f. (also called likelihood function) three methods are used to classify the image: SVM as main part of research and ML model with normal and exponential assumption to compare with SVM model. ML model with normal distribution assumption called MLC and with exponential distribution assumption called EXP. And these three models are named as: MRF-SVM, MRF-MLC; and MRF-EXP. Theory of methods is explained in chapter 3 and here implemented models are explained.

4.3.1. Prior energy

MRF

MRF model classifies image with neighborhood system Ni that considered first-order and Second-order neighbors of interested pixel consists of eight pixels. The contribution of neighbor pixels was controlled by a weighing system. Theory of these models is explained in chapter 3.

4.3.2. Conditional energy

Three adapted methods in this research for modelling conditional energy will be expressed in the following.

SVM

According to Section 3.3 SVM assigns class labels to image pixels based on decision function $f(x)$. Apply SVM model as likelihood function in MAP criterion needs to produce class probabilities instead of class labels. Class probabilities was produced based on Plott's theory (Lin, et al., 2007) where a sigmoid function is used to map the SVM outputs into probabilities. This study using the *kernlab* package implemented SVM that detailed process is on Appendix 1.

MLC

In this model, classes are assumed to follow normal distribution:

$$P(d_i \mid w_i) = \frac{1}{\sqrt{2\pi\delta_i^2}} \exp\left(- \frac{(d_i - \mu_i)^2}{2\delta_i^2} \right) \qquad (4.5)$$

Where, μ_i, δ_i^2 are mean and variance of class w_i assigned to pixel i. According to the Equation 4.5:

$$g\ (d_i \mid w_i) = -\ln P(d_i \mid w_i) = \frac{1}{2} \ln \sqrt{2\pi} + \ln \delta_i + \frac{(d_i - \mu_i)^2}{2\delta_i^2} \qquad (4.6)$$

Where a term $\frac{1}{2} \ln \sqrt{2\pi}$ is independent of d_i, w_i and can therefore be omitted.

EXP

EXP model is defined for classification of ED image. In the conditional part ML model assumes classes are exponentially distributed that is a correct assumption. Based on exponential distribution formula:

$$P(d_i \mid w_i) = \frac{1}{\gamma_i} \exp\left(-\frac{d_i}{\gamma_i}\right) \tag{4.7}$$

Where γ_i is the mean value of class w_i . Natural logarithm of Equation 4.7 is:

$$g(d_i \mid w_i) = \ln \gamma_i + \left(\frac{d_i}{\gamma_i}\right) \tag{4.8}$$

Then the likelihood energy is implemented according to decision rule of Equation 4.8.

4.3.3. Energy minimization

SA that was expressed in Section 3.2.4, minimizes the posterior energy function to label pixels and classify them into defined classes.

4.3.4. Accuracy assessment

Quality of obtained results is required to assess the performance of classification method. This accuracy assessment will define degree of assurance on the produced data for the objective application.

The most common way to express classification results is a confusion matrix. To compute a confusion matrix the results will be assessed through check the labeled data by classifier and the reference data. This $n \times n$ matrix compares classified pixels of each class with reference data of that class where, n is the number of classes.

This research evaluated obtained results using kappa coefficient (κ) according to confusion matrix. More information about kappa coefficient is provided in (Richards & Jia, 2006).

Also, the validation of obtained results considered standard deviation of kappa coefficient for 10 run. So the term reproducibility was defined to present how the classification results are confident and other performance of method can satisfy expectations.

4.3.5. Class separability

Based on distribution of classes on feature space, class probabilities are produced. These distributions often have overlaps which effect on classification of data. Due to this problem the concept of class separability is introduced. It is a indicating how well two classes are separated(Jiancheng, et al., 2010). It is a classical concept in pattern recognition and independent of coordinate system(Fukunaga, 1990). Class separability relates to distribution of classes. There are several methods to measure class separability. Richards & Jia (2006) mentioned the divergence and the Jeffries-Matusita (JM) distance to measure class separability. In this research JM was adapted for exponential distribution.

The Jeffries-Matusita (JM) distance

First $p(d \mid w_i)$ and $p(d \mid w_j)$ spectral class probability distributions are introduced that indicate the probability distribution of class i and j at the position d. Then the JM is defined as:

$$J_{ij} = \int_d \{\sqrt{p(d \mid w_i)} - \sqrt{p(d \mid w_j)}\}^2 dd \tag{4.12}$$

Which is the distance between the two class density functions. For exponentially distributed classes:

$$p(x \mid w) = \frac{1}{\gamma} \exp(-\frac{x}{\gamma}) \tag{4.13}$$

Where γ is standard deviation. Then JM becomes:

$$\mathrm{JM}_{\mathrm{ij}} = 2\left[1 - \frac{2\sqrt{\gamma_i \gamma_j}}{\gamma_i + \gamma_j}\right] \tag{4.14}$$

JM=0 indicates classes are the same and JM=2 denotes very well separable classes.

5. Results

First part of this chapter provides the results of application of SVM method without MRF part for a synthetic image described in Chapter 4. Then MRF-SVM, MRF-MLC, and MRF-EXP models were performed and results are expressed in the following.

5.1. Application of SVM

The SVM algorithm offers some flexibility to control some variables. The right selection of these variables leads to the nicer performance of SVM. In this research the effect of some variables on accuracy of SVM classification was investigated to get better result in the proposed method (MRF-SVM model). Before try different variables the size of training set was selected in Section 5.1.1, then in Section 5.1.2 model was run for both C-SVM and ν-SVM algorithms to investigate behaviour of model for different kernel functions and its parameters, after that optimum value of C and ν will be discussed in Section 5.1.3.

5.1.1. Size of training set

A training set is a portion of a data set used to train a model for prediction or classification of values. Appropriate training samples for each class helps to have a reasonable estimation of class labels. For each class, the training sample should fully describe the classes spectrally. (Richards & Jia, 2006) mentioned a recommended training set size; 10N as minimum per class and as many as 100N if it is possible, where N is dimension of multispectral space.

Like other supervised classifiers, SVM needs to be trained. SVM separates classes by identifying the support vectors from training samples. For very large number of training samples, it is sometimes impossible for SVM to use all of them to determine support vectors (Koggalage & Halgamuge, 2004). Then large training set size can limit speed of SVM. Moreover training set size can impact greatly on classification accuracy, since increasing the number of training samples leads to more accurate classification (Foody, et al., 2006). These reasons cause to this research considers the size of training set.

Since this research employs synthetic image as objective image for classification, training samples were produced through a random generator with the same class parameters of the objective image. In other words, pixels of the image were not used as training samples.

Number of training set was investigated for software default parameter setting for ED image. Both algorithms C-SVM and ν-SVM were applied whereas kernel function was Radial Basis kernel (RBF) with σ (was explained in Chapter 3) equal to 0.1 and C=1, ν=0.2. SVM Model was run 10 times in a wide range of number of training samples from 10 to 1000 (Figure 5.1). Figure 5.1a depicts accuracy

of SVM as function of training sample size and Figure 5.1b shows standard deviation of classification accuracy. Results for C-SVM show that select number of training samples equal 20 is not a wise choice, because accuracy is very small, then it can be omitted. Also it has to be notice that large number of training set (more that 1000 is very time consuming which software can not handle).

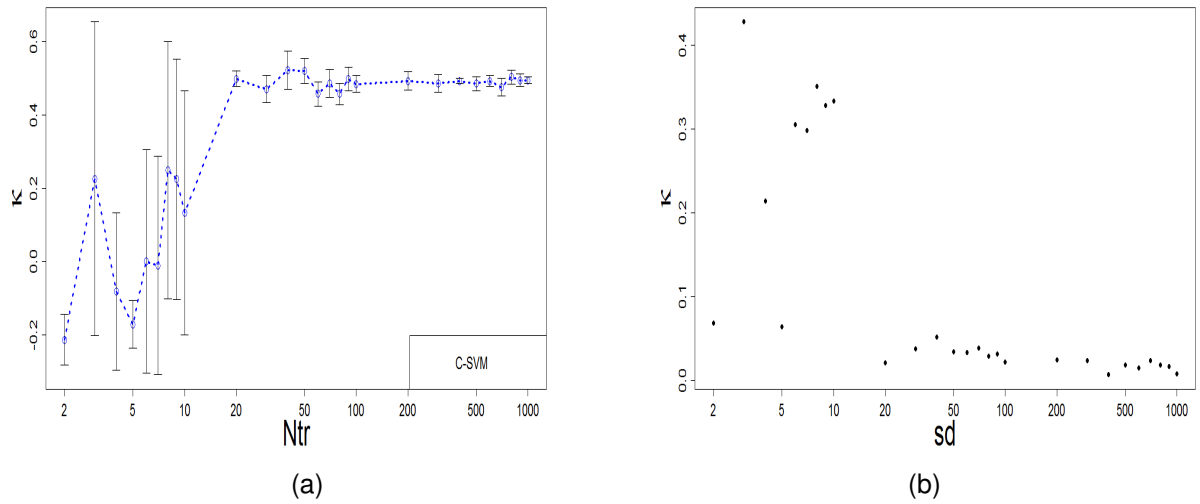


Figure 5.1: (a) Results of different number of training samples for C-SVM model where Ntr is number of training samples, (b) standard deviation of results for 10 runs of each training size where sd is standard deviation.

According to these results, range between 20 and 1000 is recommended. This choice leads to better results in accuracy and reproducibility. This research used 1000 training set because of its high accuracy and less standard deviation. Also (Foody & Mathur, 2004) study investigated size of training set and its effect on the accuracy of SVM classifier and their results showed there is a positive relation between number of training samples and accuracy of SVM. Moreover, acquire large training set size is not costly for this research while other studies may use smaller training sets.

For ν-SVM, default setting in this case did not give acceptable result, then another amount for ν was selected. The higher values ν=0.9 and ν=0.7 were tried that results are depicted in Figure 5.2a and b. Results for ν=0.7 shows number of training samples larger than 30 gives better accuracy and also less variety in results which means better reproducibility. Then range between 30 and 1000 is recommended which this research used 1000 training samples for ν-SVM too.

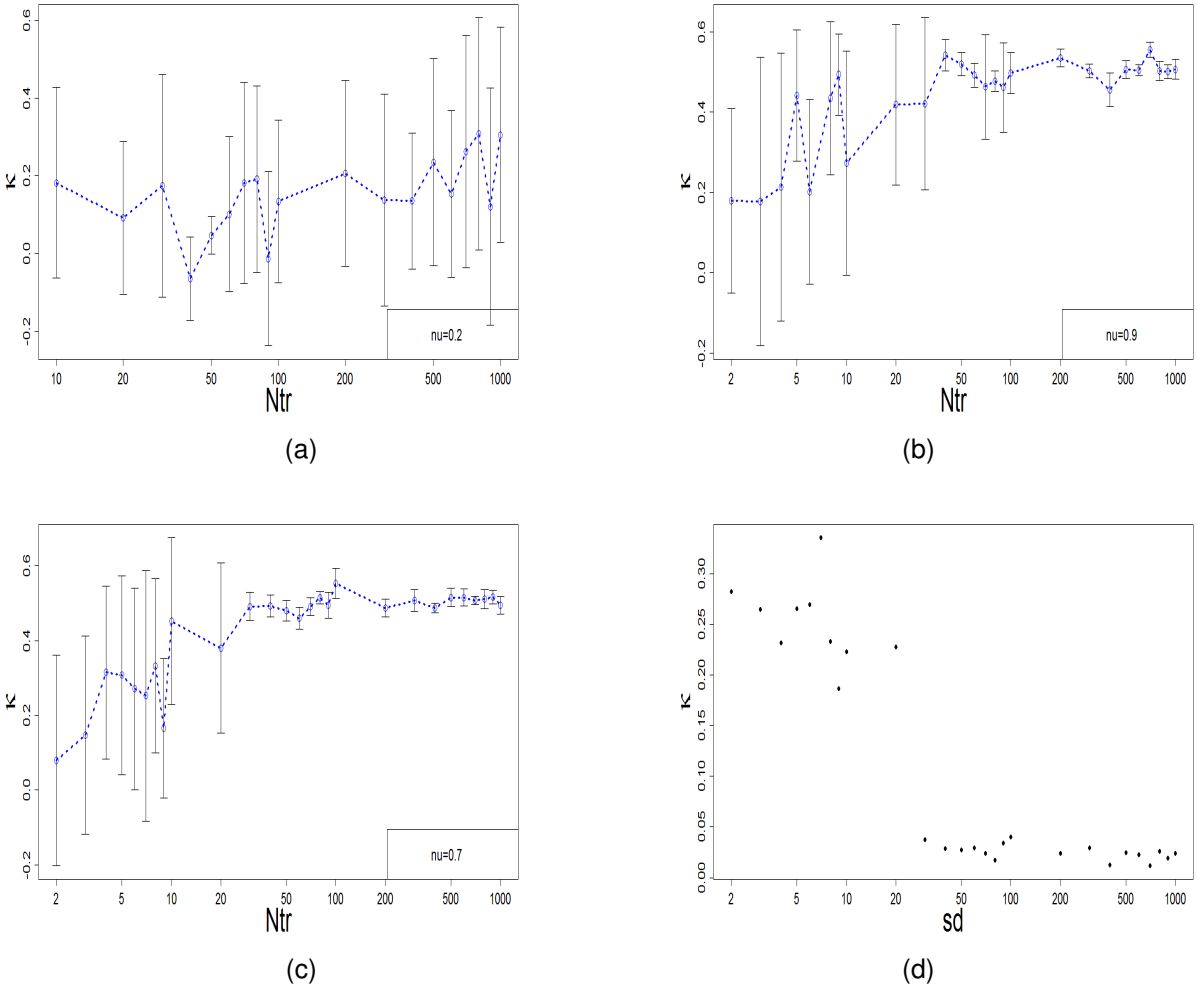


Figure 5.2: (a), (b) and (c) Different number of training samples for ν =0.2, ν =0.9 and ν =0.7 respectively where Ntr is number of training samples. (d) Standard deviation of results for 10 runs of each training size where sd is standard deviation and ν =0.7.

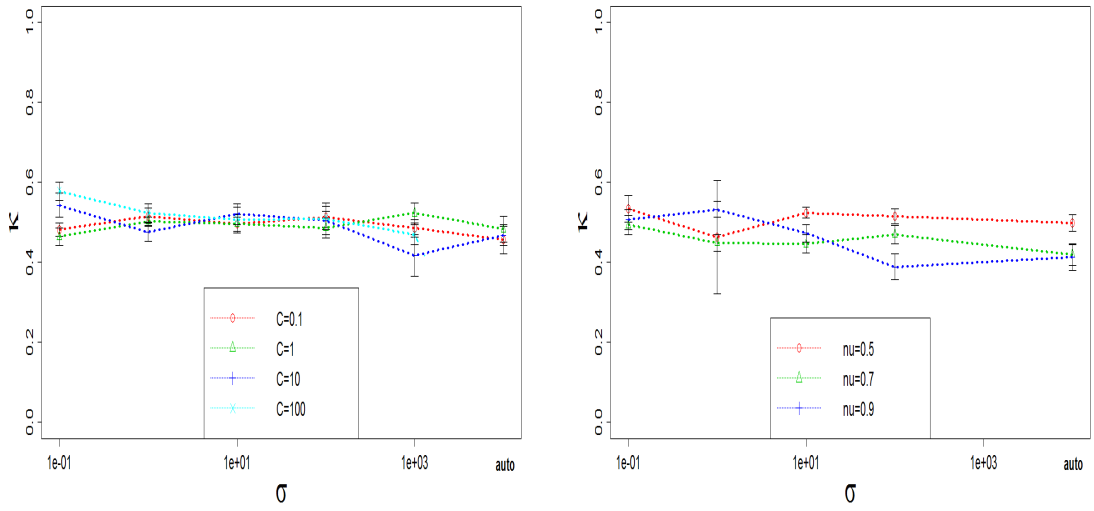
5.1.2. Kernel function

Next step is selection of kernel function. As mentioned before kernel functions can impact on performance of SVM where many studies investigated behaviour of SVM with different type of kernel functions (Huang, et al., 2002; Scholkopf, et al., 1997; Steinwart, 2002).

In non-linear cases kernels used to map input data into a high dimensional space. This transformation helps to separate data. The important part is to select appropriate kernel function to correctly map data in new space. There are several types of kernel functions which are used in remote sensing. Here we want to discuss two types of kernel functions: the Gaussian Radial Basis Function (RBF) and the linear kernel. First RBF kernel and selection of its optimum parameter is considered and then it will be compare to linear kernel function.

5.1.2.1. Radial Basis kernel Function

Radial basis function (RBF) was described in section 3.3.3. Choosing different σ can effect on Radial basis function (RBF) impact of kernel functions. This research uses different σ to investigate changes in accuracy of SVM. The value of σ is varied from 0.1 to 1000 logarithmic. Also Kernlab package in R has an option for implementing SVM which selects σ automatically from the data. Both C-SVM and ν-SVM were run by this option and also different σ for variation of C and ν. Results are shown in Figure 5.3.



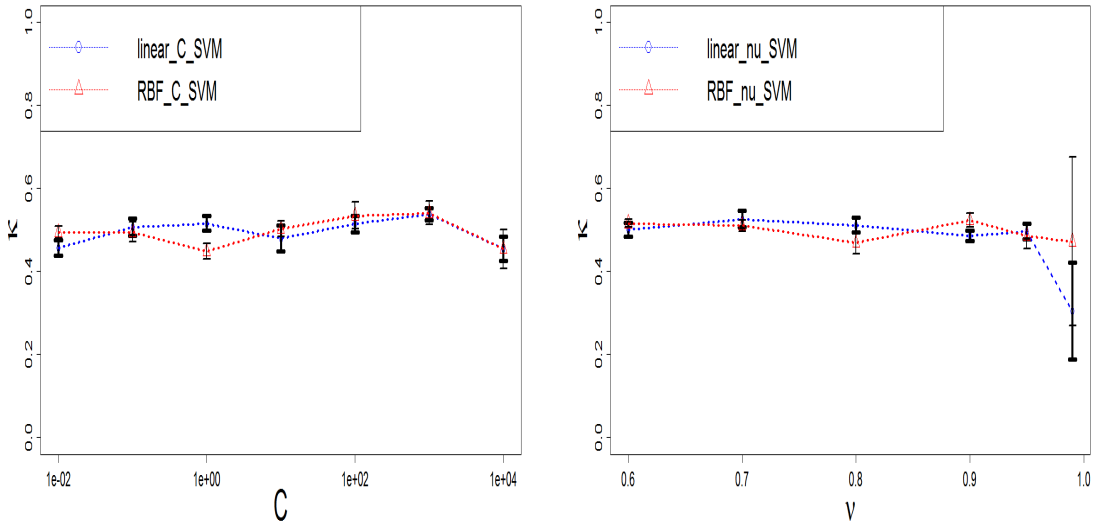
(a) (b)

Figure 5.3: (a) Different σ for C-SVM, the last one is for automatic σ selection. (b) Different σ for ν-SVM, the last one is for automatic σ selection.

These results show that there is no special effect of σ in this case for SVM accuracy, also automatic selection is not a good option for this case compared to other σ. Then the parameter σ in the following was selected as 1.

5.1.2.2. Linear kernel Function

The formulation of Linear kernel function was described in section 3.3.3. There is no parameter for this kernel. Results of implantation SVM for linear kernel function and RBF is shown in Figure 5.4 for different C and ν.



(a) (b)

Figure 5.4: compare linear kernel function and RBF kernel for (a) C-SVM and (b) ν-SVM.

Figure 5.4 shows that differences are very small which can be disregarded. For both SVM algorithms, RBF is selected to do next step.

5.1.3. C and ν

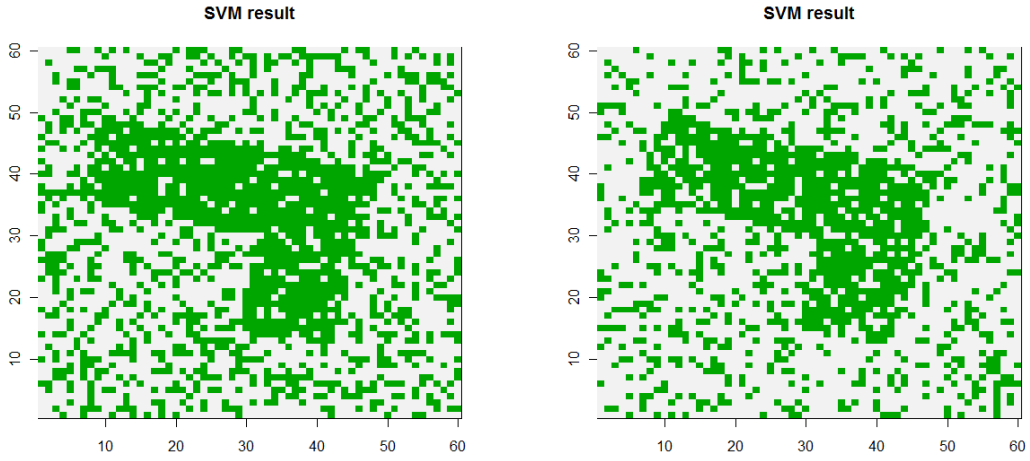
Definition of C and ν is brought in section 3.3.2 where C is penalty parameter to control misclassification samples and ν controls both errors and number of support vectors.

According to results in sections 5.1.2.1 and 5.1.2.2 different values of C and ν were selected to compare kernel functions, which these results can be considered to select optimum C and ν Figure 5.4. Note that ν-SVM can not take ν values less than 0.5 that may occur in some circumstances, more information is provided in (Perez-Cruz, et al., 2003).

These results show obtained accuracy is not too sensitive to value of C and ν, even both SVM types (C-SVM and ν-SVM) do not show any considerable difference in accuracy. Because of these C=10 is selected throughout the thesis.

5.1.4. Summary

Based on obtained results, SVM model will be implemented with C=10 and RBF kernel with δ=1. Also it will use 1000 training set to classify the image. The assumption for the following is that these variables are the same for ND image. Figure 5.5 displays classification results of SVM for selected parameters.



(a) (b)

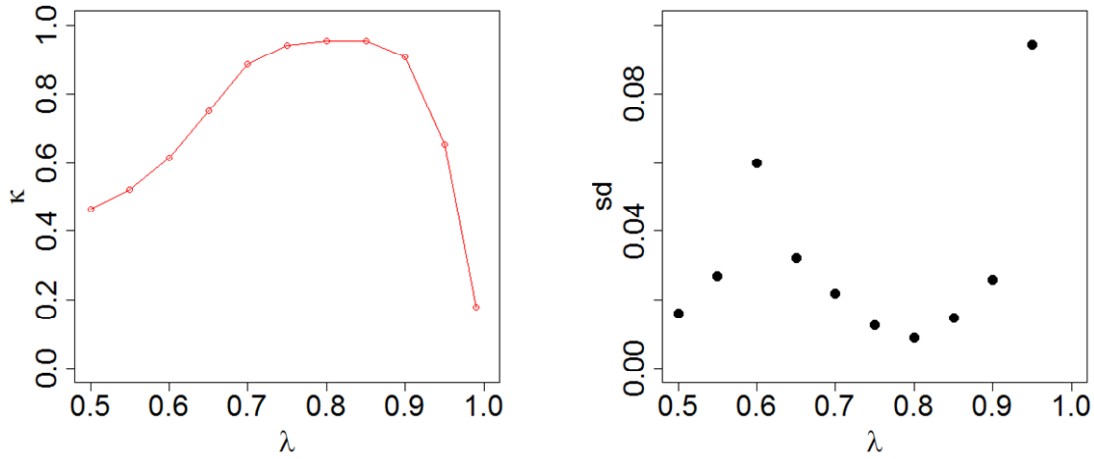
Figure 5.5: classified image with SVM where C=10 and σ =1 was used (a) ED image, (b) ND image.

5.2. Classification of ED image

5.2.1. MRF-SVM

According to results of Section 5.1, first experiment was performed MRF-SVM model 10 times for different value of parameter λ to classify ED image (
Figure **5.6**a). The study chose $\lambda = 0.5,0.55,...,0.95,0.99$. The obtained results presents a bell shaped curve that indicates the effect of λ value on the achieved κ. The maximum observed value for $\kappa_{\max}$ is 0.95 where λ=0.8 and λ=0.85. Due to the steps of 0.05 for λ, identification of optimal λ value is not very precise, then an optimal range for λ value can be defined that κ is close to $\kappa_{\max}$. Moreover, this range can be considered as a range where model is less sensitive to λ value. The criterion to choose this range was defined as $\kappa \geq 0.85$, then optimal range of λ value was between 0.7 and 0.9. In addition, standard deviation of obtained κ values was considered, which is displayed in Figure 5.6. It can be observed that for larger values of κ standard deviation is small.

Next, Table 5.1 compares results of new model with SVM model solely. It is obvious that there is a remarkable improvement in accuracy of results. Notice that SVM results are independent of λ that is.



(a) (b)

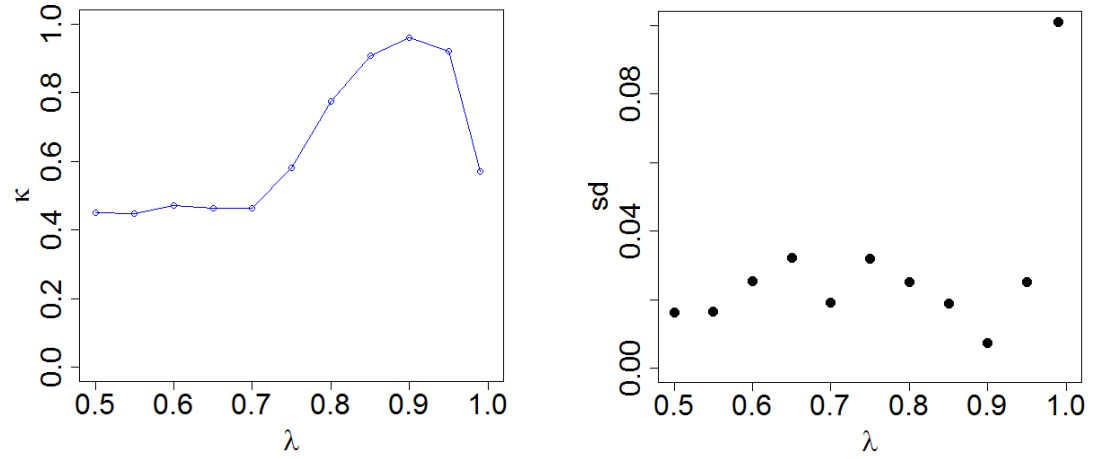
Figure 5.6: Classification results of MRF-SVM model for exponential distribution classes with different smoothness parameter λ and 10 run. (a) Accuracy of model. (b) Standard deviation of 10 runs.

	λ	κ_{max}
SVM	-	0.41 ± 0.05
MRF-SVM	0.7-<u>0.8</u>-0.9	0.95 ± 0.03

Table 5.1: Results of C-SVM model and MRF model based on C-SVM for ED image. The underlined value is optimum λ.

5.2.2. MRF-MLC

This section shows the results of MRF model where conditional model is MLC (Equation 4.6). The conditional model (MLC) assumes classes follow normal distribution while they are exponentially distributed. This part investigates how this wrong assumption effect on classification accuracy.



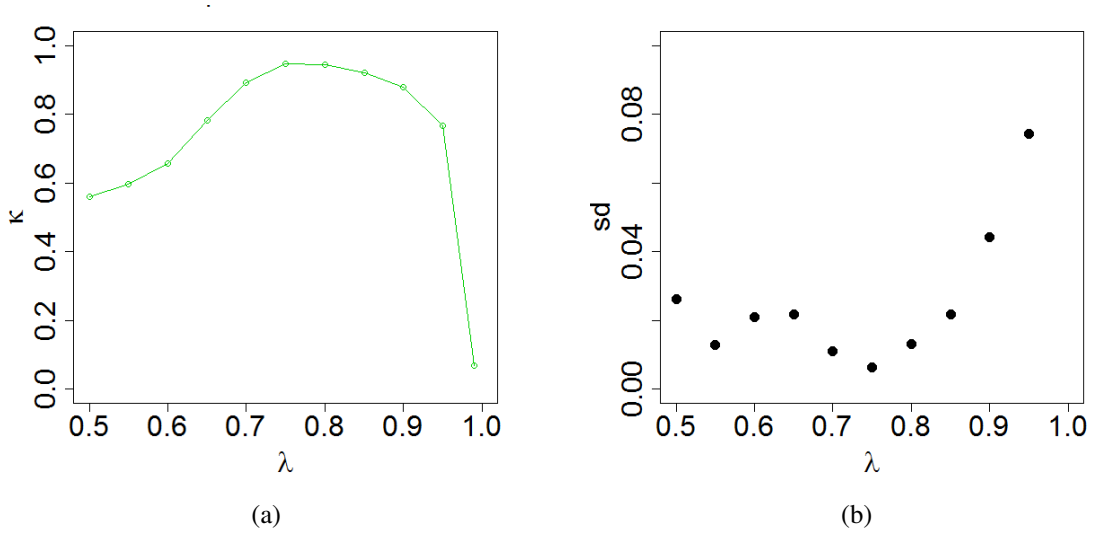
(a) (b)

Figure 5.7: Classification results of MRF-MLC model for exponential distribution classes with different smoothness parameter λ and 10 run. (a) Accuracy of results. (b) Standard deviation of results for 10 runs.

From the experiment results of this method, it can be observed that MRF model still can improve the results but requires a careful selection of λ (Figure 5.7). Figure 5.7b displays standard deviations of results are acceptable (less than 0.03) where repetition of procedure does not strongly affected the results.

5.2.3. MRF-EXP

This model aims to examine how MRF-MLC behaves when class distribution is known. In fact this part is an ideal design of conditional model using maximum likelihood with assumption of exponential distribution. Figure 5.8 a and b shows the accuracy of implemented method and its standard deviation for different λ and 10 runs. Figure 5.8a shows a high accuracy range of λ, where for λ between 0.7 and 0.9 the κ value is more than 0.8. Consider this high accuracy; variations of results are also acceptable (Figure 5.8). It means MRF-EXP model results are reliable enough (less than 0.03). In fact, obtained κ value reaches the highest accuracy that can be produced.

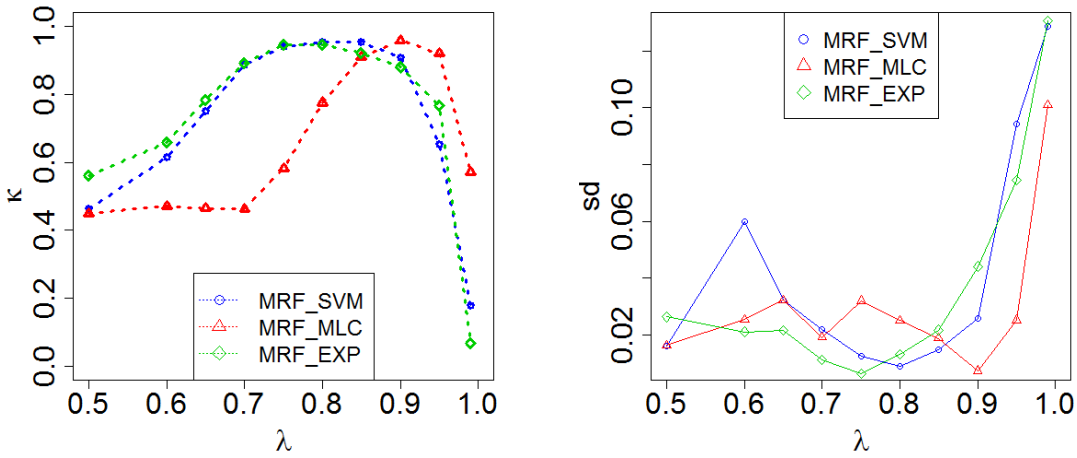


(a) (b)

Figure 5.8: Classification results of MRF-EXP model for exponential distribution classes with different smoothness parameter λ and 10 run. (a) Accuracy of results. (b) Standard deviation for 10 runs.

5.2.4. Performance of MRF-SVM, MRF-MLC, and MRF-EXP on ED image

Here the results of three models are compared in one plot (Figure 5.9), from this plot it can be seen that MRF-SVM (new model) gives accuracy as high as MRF-EXP which is ideal model for MRF based on maximum likelihood. These results show MRF-SVM even has higher accuracy for some value of λ like 0.85 and 0.9. For MRF-MLC maximum κ shifts to larger value of λ and obtaining sufficient accuracy is very sensitive to λ while the other models do not present this sensitivity.



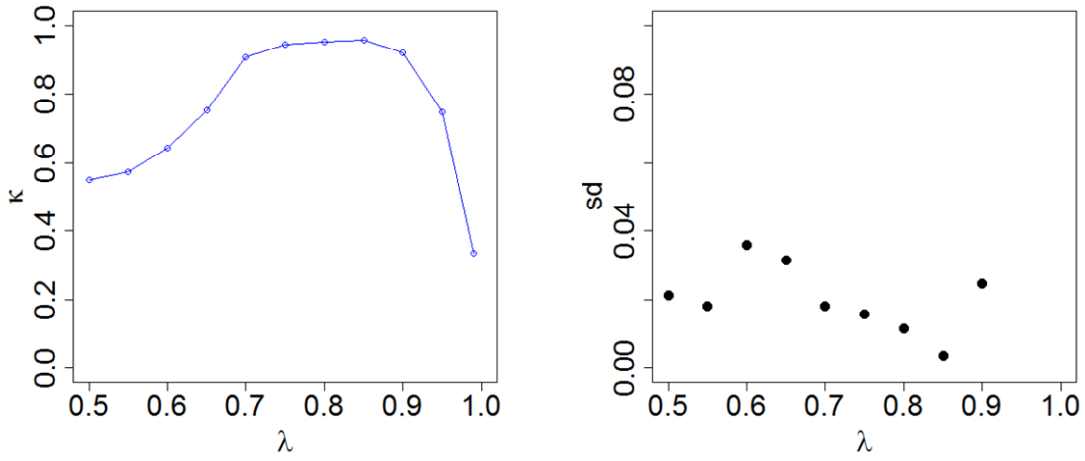
(a) (b)

Figure 5.9: Compare results of three models: MRF-SVM, MRF-MLC, and MRF-EXP.

5.3. Classification of ND image

5.3.1. MRF-SVM

MRF-SVM model in this section is applied on image with normally distributed classes. The conditional model is SVM, thus no assumption about class distributions. Parameters of the model are the same as MRF-SVM for exponential distribution classes. Model was run ten times and effect of different λ on accuracy and standard deviation of results is displayed in Figure 5.10. From these results, it can be observed that MRF-SVM gives sufficient results for ND image like ED image. The maximum observed value of $\kappa_{\max}$ is equal to 0.96 for $\lambda = 0.85$. Also there is a range of parameter λ like ED image that model's accuracy is higher than 0.8. For this range variation of results is less than 0.02. This low standard deviation is a reason for acceptable performance of MRF-SVM model. High accuracy for values of parameter λ like 0.7 means silmilar contribution of both MRF and SVM models make the new model stronger.

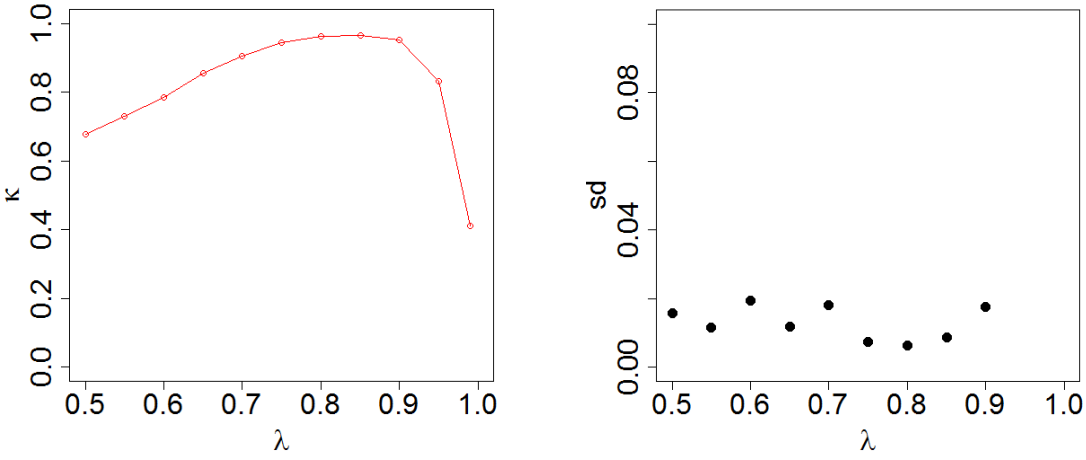


(a) (b)

Figure 5.10: κ values of MRF-SVM model for normal distribution classes with different smoothness parameter λ and 10 run. (a) κ. (b) Standard deviation of results for 10 runs.

5.3.2. MRF-MLC

For the MRF-MLC model classification of ND image is based on MLC model as conditional part. MLC assumes distribution of classes is normal that is a correct assumption and it can be said that this is equal to MRF-EXP model in exponential distribution classes. Figure 5.11 shows the results for λ between range 0.65 and 0.9 is higher than 0.8 with low standard deviation. This range for λ is wide enough that it can be said κ is not too sensitive to value of λ and also both MRF and MLC models have same contribution at the final decision to label classes. This means the role of MRF model to get high accuracy is smaller. Figure 5.11 present a similar behaviour as results of (Tolpekin & Stein, 2009) for scale=1.

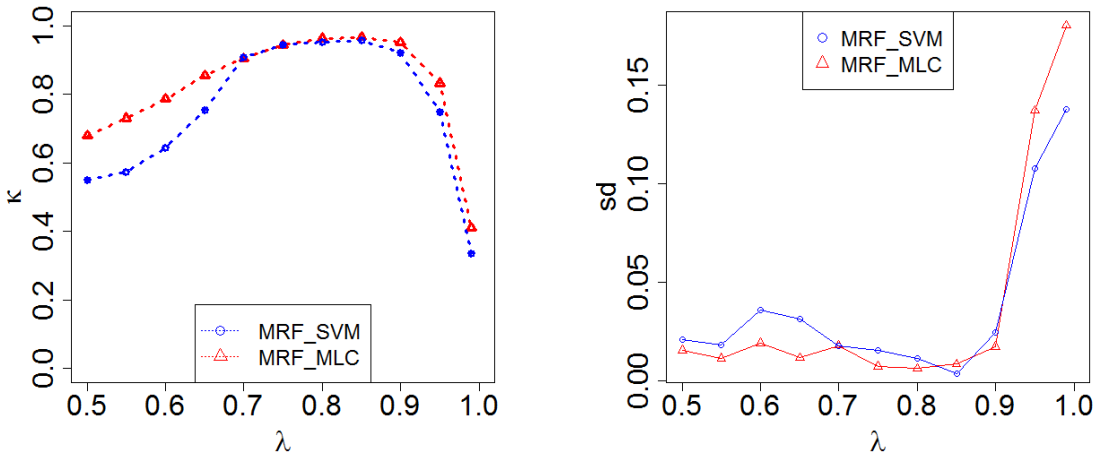


(a) (b)

Figure 5.11: κ values of MRF-MLC model for normal distribution classes with different smoothness parameter λ and 10 run. (a) κ. (b) Standard deviation of results for 10 runs.

5.3.3. Compare performance of two models for ND image

To compare two models obtained accuracy of both models were depicted in one plot (Figure 5.12). It can be seen that MRF-MLC model gives higher accuracy than MRF-SVM but two models have the similar optimal range for λ values which indicates the similar sensitivity of models to selection proper λ. In other words, if a confident range for λ is defined both models obtain similar classification accuracy. This range is between 0.7 and 0.9 where classification accuracy is close to κ_{max}. In this range standard deviation of results are also small. It shows MRF-SVM is as accurate as MRF-MLC model.

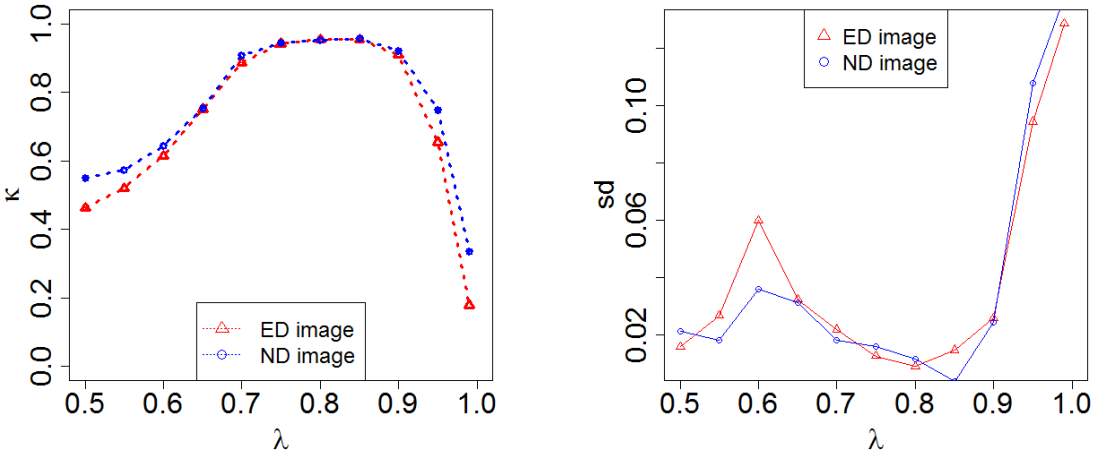


(a) (b)

Figure 5.12: Compare results of MRF-SVM and MRF-MLC (a) κ values, (b) standard deviation of κ.

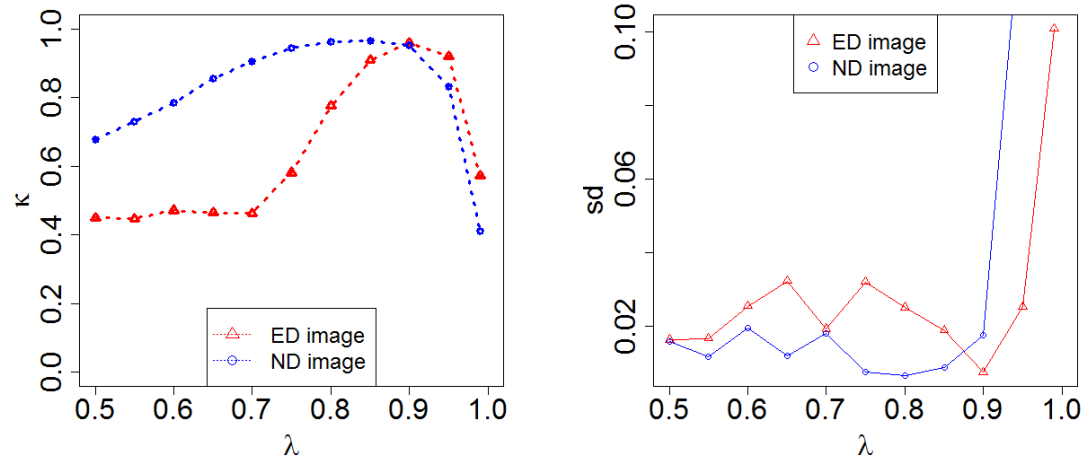
5.4. Compare performance of MRF-SVM and MRF-MLC for two images

In this section result of MRF-SVM model for both ED and ND images is plotted to investigate the behaviour of model for different class distribution (Figure 5.13). The striking point is similar behaviour of MRF-SVM for both images. Performance of the model for two images indicates similar sensitivity of model to λ value. Figure 5.13 presents MRF-SVM model for both distributions gives close κ values, it is not just about accuracies; even standard deviations of results are very close. Strength of MRF-SVM model can be seen if its results are compared to results of MRF-MLC (Figure **5.13** and Figure 5.14). MRF-MLC model has different accuracy and standard deviation for two images because its conditional model is based on distribution assumption. Figure 5.14 presents different sensitivity of model to λ value for two images where the MRF-MLC model in case of ED image has a small optimal range for λ value.



(a) (b)

Figure 5.13: MRF-SVM results for ED and ND images, (a) κ values, (b) standard deviation of κ.

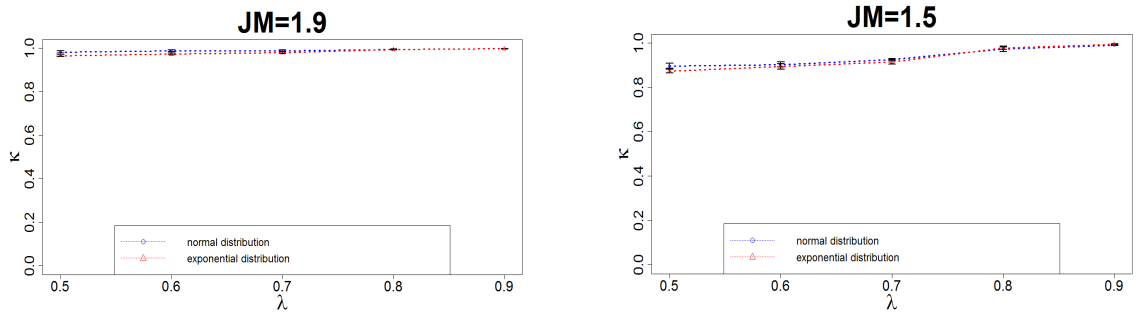


(a) (b)

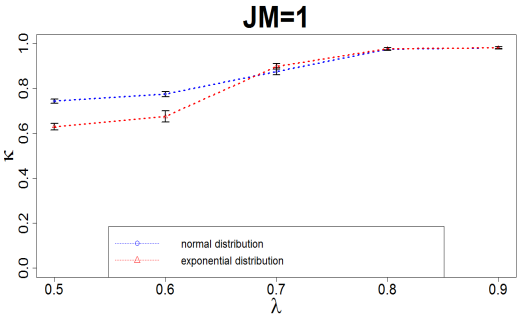
Figure 5.14: MRF-MLC results for ED and ND images, (a) κ values, (b) standard deviation of κ.

5.5. Effect of class separability

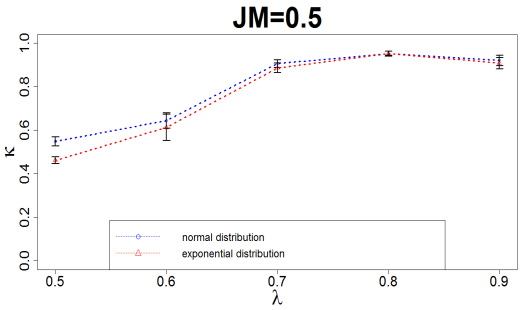
In previous sections the effect of smoothness parameter λ on results was discussed. In this section variation of class separability will be considered. Here different values for class separability are selected from JM=0.1 (poorly separability) to JM=1.9 (excellent separability). see section 4.3.5 for more details. The results are displayed for ED and ND images in Figure 5.15.



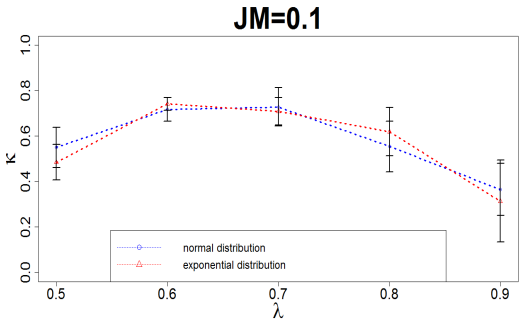
(a) (b)



(c)



(d)



(e)

Figure 5.15: Results of different class separability on MRF-SVM classification accuracy for various JM distance as a function of λ.

Experimental results from Figure 5.15 show that optimal value for parameter λ depends on class separability. According to expectation increasing class separability improves the classification accuracy. Also results of this experiment show for higher JM values (well separable classes) classification accuracy improved and sensitivity to λ value reduced. Whereas for lower class separability, the role of smoothness parameter λ becomes bigger and for different value of λ the accuracy is changed. However, experimental results shows for all values of JM the optimum smoothness parameter λ is 0.9 except of JM=0.1 where λ is 0.6 (Figure 5.15 e).

Another observation from these results is that the MRF-SVM model gives similar results for the two images especially the optimum λ value for all JM values is the same. It shows stability of model for different class distribution.

5.6. Computation time

In this section, performance speed of MRF-SVM model is illustrated in Table 5.2 .This time computation is based on number of iteration for energy minimization.

Table 5.2: Number of iteration for MRF-SVM and MRF-MLC models.

		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	0.99
ND image	MRF-SVM	66	66	76	65	63	55	52	58	52	62	76
	MRF-MLC	69	67	68	68	63	56	55	51	55	62	77
ED image	MRF-SVM	70	73	69	67	62	53	56	51	53	62	77
	MRF-MLC	61	62	62	66	67	65	66	55	52	50	60

From these results we see that number of iteration for both models is almost the same. Also MRF-SVM has similar number of iteration for ND and ED images.

5.7. Implementation on real image

To test applicability of the model for real images, it was implemented on the Envisat ASAR (C-band) satellite image in Single Look Complex (SLC) format. The image has an azimuth (along track) resolution of 4 to 5 m and range (across track) resolution of 9 to 18 m. A subset of image was selected which contains 145 × 150 pixels (Figure 5.16).

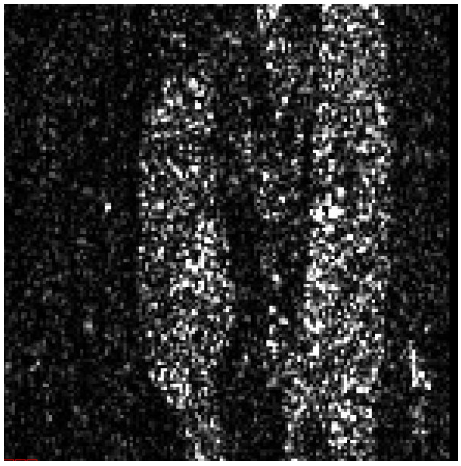
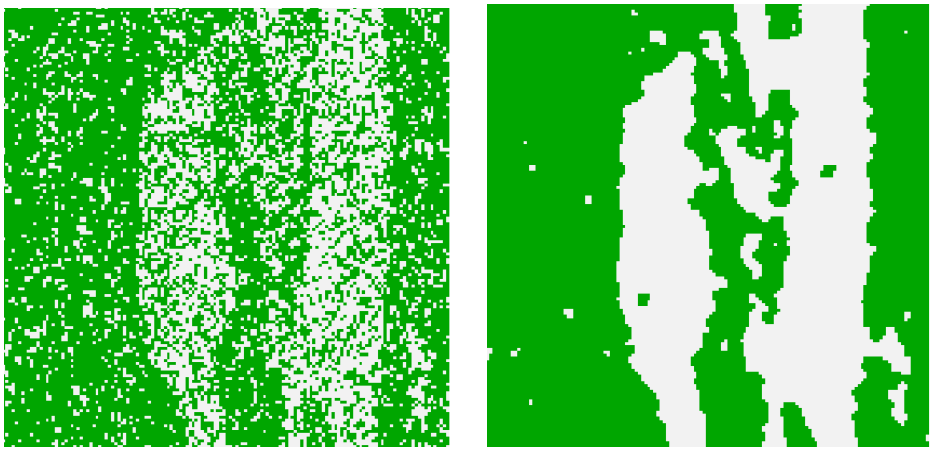
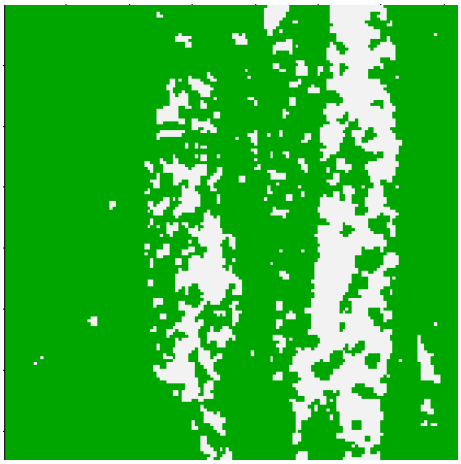


Figure 5.16: Envisat ASAR, in Single Look Complex (SLC) format with 145 × 150 pixels.



(a) (b)



(c)

Figure 5.17: (a) SVM classified image. (b) MRF-SVM classified image with λ=0.85. (c) MRF-MLC classification results with λ=0.95.

Figure 5.17a presents classified images using SVM which looks noisy and
Figure **5.17**b shows MRF-SVM results for the same image. Visual interpretation of images shows remarkable improvement for MRF-SVM classification results compared to SVM results. From the classified images it can be observed that the MRF-SVM classified image is smoother than SVM ones. Also,
Figure **5.17**c shows classified image produced by MRF-MLC. MRF-MLC can not classify image properly with respect to results of MRF-SVM where objects does not determined suficiently.

6. Discussion

This thesis introduced a new method based on MAP criteria for image classification in remote sensing. The proposed method applies SVM to model class conditional p.d.f. and employs MRF for modelling the prior p.d.f. The proposed model allows the user classify image data when no information about class distributions is available. A smoothness parameter λ was used to control balance between prior and conditional p.d.f. terms.

Application of SVM was considered by setting number of training samples. Results show κ value of both C-SVM and ν-SVM for number of training set equal to 30 is significantly improves compared to smaller number of training sets. Increasing number of training samples decreases sensitivity of classification accuracy of the model. Also standard deviation of results decreases which is expected. Due to using random generator in our case to produce training samples, the research chose 1000 number of training samples. While study of Foody & Mathur (2004) showed SVM can perform well for carefully selected small number of samples. In terms of user-defined parameters for SVM, several studies mentioned the influence of these parameters on performance of SVM (Belousov, et al., 2002; Huang, et al., 2002; Pal & Mathur, 2005; Scholkopf, et al., 1997). Our case did not present considerable sensitivity to these parameters. But the sensitivity of SVM to these parameters should be considered.

The performance of proposed method was tested on two synthetic images; ED image with exponentially distributed classes and ND image with normally distributed classes. The choice of synthetic image allowed controlling class distributions and class separability. Moreover, applying it allowed the model to use multiple input images with the same characteristics. This application of several input images indicates that obtained results are not limited to one image and the low values of standard deviations prove reproducibility of the results.

Through using a group of synthetic images model was run ten times for the same class parameters, but different pixel DN values. Different performance of the model with same parameters helps to get more reliable results whereas it indicates validity of results is not restricted to one image realization. Selection of more than ten runs may improve reliability of the results but limitation of time was considered.

Experimental results for ED image show acceptable performance of the new MRF-SVM model. The obtained results for MRF-SVM compared to standard SVM are improved significantly as Table 5.1 illustrates. This was expected while the new model uses more information of input data. It can be explained based on MAP criteria that benefits from MRF. One of the difficulties in using MAP criteria is the lack of prior information about input data. Application of MRF can overcome this problem and improve the results. It should be noticed that classified image includes is smooth, which influences on the accuracy of MRF model positively. Also the MRF-SVM model' results show a range for λ value where the accuracy is close to maximum. Similar range is observed from results of MRF-EXP model.

It may occur due to suitable conditional models where the results of MRF-MLC do not show such wide optimal for range of λ (Figure 5.10). From these results it can be inferred that, SVM model has the same effect as EXP on the accuracy of classification whereas SVM does not require knowledge of class distributions. This independency of SVM to class distribution information is a strength point for introduced model. Results show for maximum accuracies reproducibility of the model is acceptable too. This is another reason that makes the model attractive.

Similarly to ED image, MRF-SVM model was implemented on ND image. Results of new test present improved accuracy respect to SVM model (Table 5.1). It can be observed from Figure 5.10 that the proposed model leads to high accuracy with good reproducibility. Comparison of the model with MRF-MLC as ideal MRF based model for ND image presented similar behaviour (Figure 5.12). MRF-MLC for some λ values gives better results but the maximum classification accuracy for both models is similar. Better performance of the MRF-MLC model is attributed to its correct assumption on data distribution for ND image.

In the case of comparison the results of new model for two images, we see a stable performance of MRF-SVM model (Figure 5.13), whereas there is no similarity for MRF-MLC model. In fact, by this comparison the influence of conditional model is considered. In MRF-MLC model, the likelihood part makes assumption based on class distribution. If this assumption is not correct or similar to the input data it may affect on accuracy of classification for different type of data (Figure 5.14). In spite of MRF-MLC or better to say MRF-ML with two different distribution assumptions, MRF-SVM model shows similar productions for two images. It indicates that SVM model can incorporate with MRF for different input data.

The study also quantified the effect of class separability (was described in Section Class separability) on the accuracy of the proposed method. Trend of model for both images is similar (Figure 5.15). For small values of JM that denotes big overlap for class distributions, role of prior model is not similar to large JM values. In other words, for big overlap of classes except of lower accuracy the optimum value of parameter λ becomes smaller. It shows conditional energy makes better estimation of class probability than prior energy. For each JM value one optimal λ can be observed which is similar to findings of (Tolpekin & Stein, 2009). Based on their study for each class separability one optimal λ value exists that can be estimated to enhance quality of results.

In terms of computational time, the research considered number of iteration for SA. Figure 5.12 shows number of iteration for both MRF-SVM and MRF-MLC model. It indicates two models require the same amount of iteration.

Applicability of developed method on SAR image was illustrated in section 5.7. Visual interpretation of the image shows MRF-SVM can improve classification accuracy compared to SVM model MRF-MLC.

7. Conclusion and recommendations

7.1. Conclusion

The objective of this study was a MRF-MAP framework that applies SVM as conditional model. Before making conclusion it should be mentioned that all the research questions are properly answered. The results show that the SVMs can be satisfactorily incorporated into the MRF model and improve the accuracy of classification. The integrated model is applied in two types of synthetic images with normal and exponential distribution classes that are called ND and ED respectively. Applicability of the model was investigated in terms of class distribution and separability as a function of smoothness parameter λ.

To apply the proposed model, first application of SVM was investigated. SVM was trained using training samples produced through a random generator with the same class parameters of the objective image. Performance of SVM is affected by some user-defined parameters. This research addressed them and adjusted proper parameters for the study. This adjustment considered two available algorithms C-SVM and ν-SVM. C-SVM with a Radial Basis kernel Function (RBF) was used for implementation of SVM.

In terms of class distribution, the maximum κ for ND and ED images is 0.96 and 0.95 with the same λ value. Results of MRF-SVM classification for both images were almost identical. Obtained results were compared with MRF-MLC and MRF-EXP models. For ND image, classification accuracy of the new model is comparable with MRF-MLC with similar optimal range of λ. But for ED image the introduced model gives better result compared to MRF-MLC. In this case MRF-SVM introduced an extensive range for λ with respect to MRF-MLC model. Performance of MRF-SVM was compared to results of MRF-EXP which using true data distribution for ED image. In this case also model has comparable results.

Investigation the effect of class separability on the accuracy of MRF-SVM presents similar behaviour of the model for two images. From the results it can be concluded that there is a relation between class separability and the optimal λ value. Results show that the optimum value of λ depends on the separability of classes. For poorly separable classes optimum λ and accuracy of classification are smaller than well separable classes. In addition, computational time of the model was compared to MRF-MLC model in terms of number of iterations. The comparison yields similar results.

The advantage of the proposed model is that: modelling conditional probability through SVM makes no assumption about class distribution in contrast to MLC. Then the model needs no information about class distribution that empirical results prove it by high classification accuracy achieved for different class distributions.

It can be concluded from the results that MRF based on SVM model is applicable in remote sensing image classification. The model performed well and produced sufficient classification accuracy in different circumstances. An optimal range of smoothness parameter value exists for the introduced model for which classification accuracy is close to optimal. Even for poorly separable classes use proper value of this smoothness parameter can enhance the results.

7.2. Recommendations

The model introduced in this study appears as a promising technique. Undoubtedly, the model still needs further development. To address the limitations, the following is recommended for further research:

1. Due to time limitations, all the characteristic of image data were not surveyed during the study. Therefore, the shape of object can be considered to study its effect on the results. Also, an increasing the number of objects to investigate how results will be affected is suggested.

2. Applicability of the model for SAR images was illustrated in this research. Performance of the MRF-SVM on SAR or QuickBird images can be studied to investigate its capability.

3. Performance of the SVM model is affected by user-defined parameters, such as kernel function and its parameters (Huang, et al., 2002; Pal & Mathur, 2005; Tso & Mather, 2009). This research adjusts SVM parameters experimentally. In further steps to improve MRF-SVM model, adjustment methodologies like grid search or gradient descent method can be considered.

4. Since SA is a time consuming approach for energy minimization due to its cooling schedule (Tso & Mather, 2009), other algorithms maybe used instead. An alternative is use of the Graph base algorithms, which its application to MRF models produces good results (Karimov, 2010).

References

Achlioptas, D., McSherry, F. , & Scholkopf, B. . (2002). Sampling techniques for kernel methods. *Advances in Neural Information Processing*

Systems 14, 335-342.

Barker, S., & Rayner, P. (1997). Unsupervised image segmentation using Markov Random Field models. In M. Pelillo & E. Hancock (Eds.), *Energy Minimization Methods in Computer Vision and Pattern Recognition* (Vol. 1223, pp. 165-178): Springer Berlin / Heidelberg.

Belousov, A. I., Verzakov, S. A., & von Frese, J. (2002). A flexible classification approach with optimal generalisation performance: support vector machines. [doi: DOI: 10.1016/S0169-7439(02)00046-1]. *Chemometrics and Intelligent Laboratory Systems, 64*(1), 15-25.

Berthod, M., Kato, Z., Yu, S., & Zerubia, J. (1996). Bayesian image classification using Markov random fields. [doi: DOI: 10.1016/0262-8856(95)01072-6]. *Image and Vision Computing, 14*(4), 285-295.

Bruzzone, L., & Persello, C. (2009). A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples. *IEEE Transactions on Geoscience and Remote Sensing, 47*(7), 2142-2154.

Bruzzone, L., & Prieto, D. F. (2000). Automatic analysis of the difference image for unsupervised change detection. *Geoscience and Remote Sensing, IEEE Transactions on, 38*(3), 1171-1182.

Burges, C.J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. [10.1023/A:1009715923555]. *Data Mining and Knowledge Discovery, 2*(2), 121-167.

Camps-Valls, G., & Bruzzone, L. (2005). Kernel-based methods for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on, 43*(6), 1351-1362.

Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks.* Paper presented at the Machine Learning.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines : and other kernel-based learning methods*: Cambridge University Press.

Fjortoft, R., Delignon, Y., Pieczynski, W., Sigelle, M., & Tupin, F. (2003). Unsupervised classification of radar images using hidden Markov chains and hidden Markov random fields. *Geoscience and Remote Sensing, IEEE Transactions on, 41*(3), 675-686.

Foody, G.M., & Mathur, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. [doi: DOI: 10.1016/j.rse.2004.06.017]. *Remote Sensing of Environment, 93*(1-2), 107-117.

Foody, G.M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. [doi: DOI: 10.1016/j.rse.2006.03.004]. *Remote Sensing of Environment, 104*(1), 1-14.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition (2nd ed.)*: Academic Press Professional, Inc.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-6*(6), 721-741.

Gualtieri, J. A., & Cromp, R. F. (1998). Support Vector Machines for Hyperspectral Remote Sensing Classification. *In Proceedings of the 27th AIPR Workshop:Advances in Computer Assisted Recognition*, 221-232.

Hermes, L., Frieauff, D., Puzicha, J., & Buhmann, J. M. (1999). *Support vector machines for land usage classification in Landsat TM imagery.* Paper presented at the Geoscience and Remote Sensing Symposium, 1999. IGARSS '99 Proceedings. IEEE 1999 International.

Hu, R., & Fahmy, M. M. (1991). *Texture segmentation based on a hierarchical Markov random field model.* Paper presented at the Circuits and Systems, 1991., IEEE International Sympoisum on.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*(4), 725 - 749.

Jiancheng, Sun, Chongxun, Zheng, Xiaohe, Li, & Yatong, Zhou. (2010). Analysis of the Distance Between Two Classes for Tuning SVM Hyperparameters. *Neural Networks, IEEE Transactions on, 21*(2), 305-318.

Karatzoglou, A., Meyer, D , & Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software, 15*(9), 1-28.

Karatzoglou, A., Smola, Alexandros, Hornik, Kurt, Zeileis, Achim, , , & (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software, 11*(i09).

Karimov, A. (2010). *Graph cuts for fast optimization in Markov random field based remote sensing image analysis.* University of Twente Faculty of Geo-Information and Earth Observation ITC, Enschede.

Kasetkasem, T., Arora, M.K., & Varshney, P. K. (2005). Super-resolution land cover mapping using a Markov random field based approach. [doi: DOI: 10.1016/j.rse.2005.02.006]. *Remote Sensing of Environment, 96*(3-4), 302-314.

Koggalage, R. , & Halgamuge, S. (2004). Reducing the Number of Training Samples for Fast Support Vector Machine Classification. *Neural Information Processing, Vol. 2, No.3*, 56-65.

Li, S.., Wang, H., Chan, K., & Petrou, M. (1997). Minimization of MRF Energy with Relaxation Labeling. *Journal of Mathematical Imaging and Vision, 7*(2), 149-161.

Li, Stan. (2009). *Markov Random Field Modeling in Image Analysis (Advances in Pattern Recognition)*: Springer.

Lin, H., Lin, C.J., & Weng, R. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning, 68*(3), 267-276.

Magnussen, S., Boudewyn, P., & Wulder, M. (2004). Contextual classification of Landsat TM images to forest inventory cover types. *International Journal of Remote Sensing, 25*(12), 2421 - 2440.

Melgani, F., & Bruzzone, L. (2002). *Support vector machines for classification of hyperspectral remote-sensing images.* Paper presented at the Geoscience and Remote Sensing Symposium, 2002. IGARSS '02. 2002 IEEE International.

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *Geoscience and Remote Sensing, IEEE Transactions on, 42*(8), 1778-1790.

Melgani, F., & Serpico, S. B. (2003). A Markov random field approach to spatio-temporal contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing, 41*(11), 2478-2487.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics, 21*(6), 1087-1092.

Pal, M. (2006). Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. *International Journal of Remote Sensing, 27*(14), 2877-2894.

Pal, M., & Mathur, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing, 26*(5), 1007 - 1011.

Perez-Cruz, F., Weston, J., Herrmann, D.J.L., & Schölkopf, B. (2003). Extension of the nu-SVM range for classification. In J. Suykens, G. Horvath, S. Basu, C. Micchelli & J. Vandewalle (Eds.), *Advances in Learning Theory: Methods, Models and Applications* (Vol. 190, pp. 179-196). Amesterdam: IOS Press.

Richards, J.A., & Jia, Xiuping. (2006). *Remote sensing digital image analysis : an introduction* (Fourth edition ed.). Berlin etc.: Springer-Verlag.

Sarkar, A., Biswas, M. K., Kartikeyan, B., Kumar, V., Majumder, K. L., & Pal, D. K. (2002). A MRF model-based segmentation approach to classification for multispectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on, 40*(5), 1102-1113.

Sarkar, A., Biswas, M. K., & Sharma, K. M. S. (2000). A simple unsupervised MRF model based image segmentation approach. *Image Processing, IEEE Transactions on, 9*(5), 801-812.

Scholkopf, B., Kah-Kay, S., Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *Signal Processing, IEEE Transactions on, 45*(11), 2758-2765.

Schölkopf, B., & Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*: The MIT Press.

Schölkopf, B., Smola, A., Williamson, R.., & Bartlett, P. (2000). New Support Vector Algorithms. *Neural Comput., 12*(5), 1207-1245.

Solberg, A. H. S., Taxt, T., & Jain, A. K. (1996). A Markov random field model for classification of multisource satellite imagery. *Geoscience and Remote Sensing, IEEE Transactions on, 34*(1), 100-113.

Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res., 2*, 67-93.

Tolpekin, V. A., & Stein, A. (2009). Quantification of the Effects of Land-Cover-Class Spectral Separability on the Accuracy of Markov-Random-Field-Based Superresolution Mapping. *Geoscience and Remote Sensing, IEEE Transactions on, 47*(9), 3283-3297.

Tso, B., & Mathur, P. M. (1999). Classification of multisource remote sensing imagery using a genetic algorithm and Markov random fields. *Geoscience and Remote Sensing, IEEE Transactions on, 37*(3), 1255-1260.

Tso, B., & Mather, P.M. (2009). *Classification methods for remotely sensed data* (Second edition ed.). Boca Raton: CRC.

Tso, B., & Olsen, R. C. (2005). A contextual classification scheme based on MRF model with improved parameter estimation and multiscale fuzzy line process. [doi: DOI: 10.1016/j.rse.2005.04.021]. *Remote Sensing of Environment, 97*(1), 127-136.

Vapnik, V. (1995). *The nature of statistical learning theory*: Springer-Verlag New York, Inc.

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data (Information Science and Statistics)*: Springer.

Waske, B., van der Linden, S., Benediktsson, J. A., Rabe, A., & Hostert, P. (2010). Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data. *Geoscience and Remote Sensing, IEEE Transactions on, 48*(7), 2880-2889.

Wei, J., & Gertner, I. (2003). MRF-MAP-MFT visual object segmentation based on motion boundary field. [doi: DOI: 10.1016/S0167-8655(03)00180-6]. *Pattern Recognition Letters, 24*(16), 3125-3139.

Xia, G. S., He, C., & Sun, H. (2007). Integration of synthetic aperture radar image segmentation method using Markov random field on region adjacency graph. *Radar, Sonar & Navigation, IET, 1*(5), 348-353.

Apendix 1

```
#*******************************************************************************
# Block 1: variable definitions, data import, preparation
#*******************************************************************************

require(MASS)
require(mvtnorm)
require(kernlab)

# Image dimensions
M <- 60
N <- 60

# Number of classes
Ncl <- 2

# How many runs
Nrun <- 10

# Here real window size is 2*WSize+1
WSize   <- 1

# Number (maximal) of pixel neighbours
Nn      <- (WSize*2+1)^2-1

x <- 1:M
y <- 1:N

# Number of bands
Nb <- 1

Path <- 'G:\\data_6oct\\exponential\\MRF_SVM_6oct\\'

dir.create(Path, recursive = TRUE)

Ref     <- array(0, c(M,N))
D       <- array(0, c(M,N))
F       <- array(0, c(M,N))
Initial <- array(0, c(M,N))
```

```
Neigh_Coord      <- array(0, c(M, N, 4))
Weight           <- array(0, c(2*WSize+1, 2*WSize+1))
mu   <- array(0,c(Ncl))                          # Mean values of Ncl classes
sigma  <-  array(0,c(Ncl))
Cinv <- array(0,c(Ncl))

#*******************************************************************************
#          Generate neighbourhood list
#*******************************************************************************
# Function assigning weights in the neighbourhood
Fw <- function(a,b){

          val <- a^2 + b^2
          val <- 1 / val
#         val <- 1 / sqrt(Camps-Valls & Bruzzone)
          val <- val^(0.5)

          val[val==Inf]<-0

          return(Camps-Valls & Bruzzone)
}

   for(k in 1:(2*WSize+1))
   for(l in 1:(2*WSize+1))
   {
          Weight[k, l] <- Fw(k-(WSize+1),l-(WSize+1))
   }

   Weight <- Weight/ sum(Weight)

for(i in 1:M)
for(j in 1:N)
{

   imin <- i - WSize
   imax <- i + WSize
   jmin <- j - WSize
   jmax <- j + WSize

   if(imin<1)  imin <-1
   if(imax>M) imax <-M
   if(jmin<1)  jmin <-1
   if(jmax>N) jmax <-N

   Neigh_Coord[i, j, ] <- c(imin,imax,jmin,jmax)
```

```
}
#*****************************************************************************
# Import Reference image
Filename <- '2classs.arr'

temp <- dget(paste(Path,Filename,sep=""))              #load image array

Ncl <- max(temp)                                       #how many colors

Ref<-temp

par(mfrow=c(1,2))
image(x,y, Ref, col=terrain.colors(Ncl), main = 'Reference',xlab='',ylab='')          #display image
plot(c(0,10),c(0,10))
Cl_colors<- terrain.colors(Ncl)
legend("right",c('Class1','Class2'),fil=terrain.colors(Ncl),cex=1.2)

#*****************************************************************************
# End of Block 1: variable definitions, data import
#*****************************************************************************
#*****************************************************************************
# Block 2: Define land cover classes infromation; generate the synthetic multispectral image.
#*****************************************************************************
#
# Class means and covariances
#
# generate random numbers from exponential distribution
rand_exp <- function(n,sigma)
{
  xt <- rnorm(n,0,sqrt(0.5*sigma))
  yt <- rnorm(n,0,sqrt(0.5*sigma))
  I <- xt^2 + yt^2
  return(I)
}

# generate random numbers from Rayleigh distribution
rand_Ral <- function(n,sigma)
{
  xt <- rnorm(n,0,sqrt(0.5*sigma))
  yt <- rnorm(n,0,sqrt(0.5*sigma))
  I <- xt^2 + yt^2
  I <- sqrt(I)
  return(I)
}

#Fix JM:
```

```
JM <- 0.5
sigma[2] <- 50.0

#temporary auxilary variable
a <- 1-JM/2

sigma[1] <- sigma[2]*(1/a-sqrt((1/a^2)-1))^2

# Check JM
i<-1
j<-2

JM  <- 2*(1-2*sqrt(sigma[i]*sigma[j])/(sigma[i]+sigma[j]))
JM

mu <- sigma

for(krun in 1:Nrun)
{

Num <- array(0,Ncl)

for(k in 1:Ncl)
{
  N0<-sum(Ref==k)

  D[Ref==k] <- rand_exp(N0,mu[k])
}

for(k in 1:Ncl)
mu[k] <- mean(D[Ref==k])

for(k in 1:Ncl)
sigma[k] <- sd(D[Ref==k])

x11()
par(mfrow=c(1,1))
image(x,y, D, col=gray((0:255)/255),cex.axis=2,cex.lab=3, main = "Image", xlab='',ylab='')

D <- round(D,digits=2)

#hist(D)

write.table(D, file =
paste(Path,'SyntheticExpo_C_SVM_lambda=',lambda,'Run=',krun,'.txt',sep=''),append=FALSE,quote=
```

```
TRUE,sep ="
",eol="\n",na="NA",dec=".",row.names=FALSE,col.names=FALSE,qmethod=c("escape","double"))


#*****************************************************************************
# End of Block 2: Define land cover classes information; generate the synthetic multispectral image.
#*****************************************************************************
#*****************************************************************************
# Block 3: SVM classification of the image D
#*****************************************************************************
# SVM training

# SVM parameters
sigma_SVM <- 1
C_SVM <- 10
nu_SVM <- 0.9

Ntrset<- 1000

Trainingset <-
data.frame(z=c(rand_exp(Ntrset,mu[1]),rand_exp(Ntrset,mu[2])),class=c(rep(1,Ntrset),rep(2,Ntrset)))

# Linear kernel
#svm_model <- ksvm(class~.,data=Trainingset,type="C-
svc",kernel="vanilladot",C=C_SVM,prob.model=TRUE)

# Radial Basis kernel "Gaissian"
svm_model <- ksvm(class~.,data=Trainingset,type="C-svc",cache =
2000,kernel="rbfdot",kpar=list(sigma=sigma_SVM),C=C_SVM,prob.model=TRUE)

# nu-SVM classification
#svm_model <- ksvm(class~.,data=Trainingset,type="nu-svc",cache =
2000,kernel="rbfdot",kpar=list(sigma=sigma_SVM),nu=nu_SVM,prob.model=TRUE)

# Apply SVM

A <- data.frame(z=as.vector(D))

SVM <- array(0,c(M,N))
CProb <- array(0,c(Ncl,M,N))
Ucond <- array(0,c(Ncl,M,N))

test <- predict(svm_model, A, type="probabilities")

for(k in 1:Ncl)
CProb[k,,]<- test[,k]
```

```
SVM[,] <- 1

SVM[CProb[1,,]<CProb[2,,]]<-2

image(x,y, SVM, main = "SVM result", col=terrain.colors(Ncl), xlab="",ylab="")

eps <- 1.0e-9

CProb[CProb==0.0] <- eps

Ucond <- -log(CProb)

#x11()
par(mfrow=c(1,2))
for(k in 1:Ncl) image(x,y,CProb[k,,],col=gray((0:255)/255))

# Accuracy assessment of SVM result

ConfSVM <- array(0, c(Ncl,Ncl))

for(i in 1:Ncl)
for(j in 1:Ncl)
{
   ConfSVM[i,j] <- sum((SVM==i)&(Ref==j))
}

ConfSVM
s1<-0
for(i in 1:Ncl)
{
  s1 <- s1 + sum(ConfSVM[i,])*sum(ConfSVM[,i])
}

kappaSVM <- (M*N*sum(diag(ConfSVM)) - s1) / ((M*N)^2 - s1)
kappaSVM

#*******************************************************************************
# End of Block 3: SVM classification of the image D
#*******************************************************************************
```