Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 06

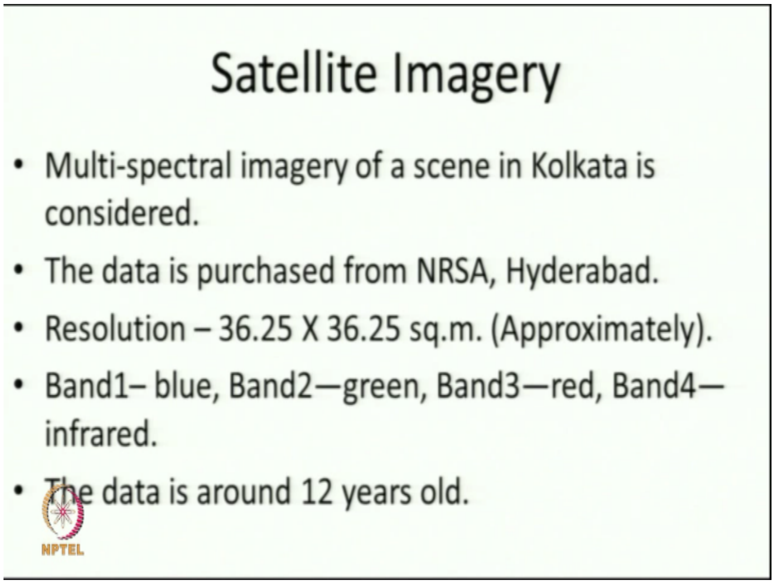
Lecture 09

Examples of Real-Life Dataset

Prof. C.A. Murthy
Machine Intelligent Unit,
Indian Statistical Institute, Kolkata

So we have been discussing quite a bit of theory of pattern recognition in several classifiers and in some of the previous classes we looked at the how the classification rules perform over artificial data sets I island I already gave you some assignments on that today I shall discuss the utility of the schemes the pattern or ignition schemes on a real-life data set.

(Refer Slide Time: 01:01)



Satellite Imagery

- Multi-spectral imagery of a scene in Kolkata is considered.
- The data is purchased from NRSA, Hyderabad.
- Resolution – 36.25 X 36.25 sq.m. (Approximately).
- Band1– blue, Band2—green, Band3—red, Band4—infrared.
- The data is around 12 years old.

So as you can see on your screen light image data for the data analysis part the satellite imagery is taken from NRSA Hyderabad as you can see from there it is 12 years old actually it is probably more than 12years old I do not remember when we bought it, it maybe even the later part of 1990s it is a scene from Calcutta it is a multispectral imagery there are four bands there and band one corresponds to blue, band 2 corresponds to green, band 3 corresponds to red and band for it corresponds to infrared.

The resolution is approximately for each pixel the resolution is approximately 36.25*36.25 square meters that is each pixel on an average it covers an area of that much that 636 those many square meters approximately this is how the image from I mean the band one image looks like.

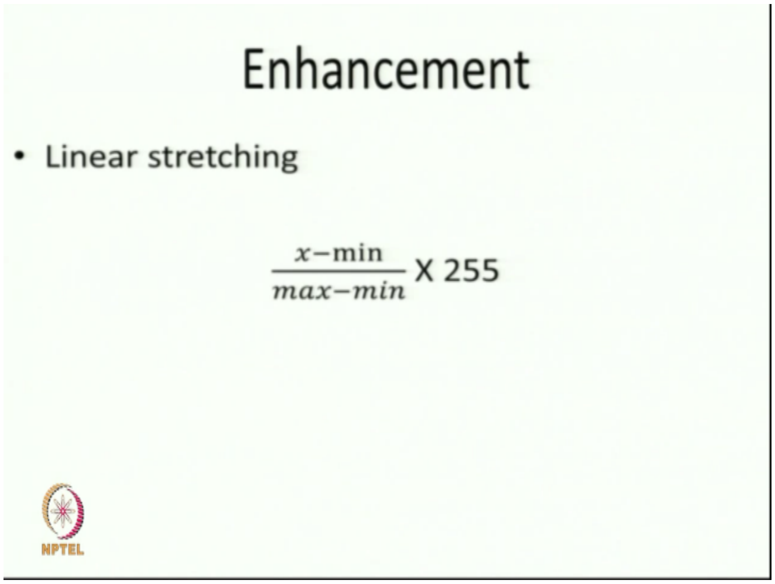
(Refer Slide Time: 02:12)

Band1 image

As you can see it is really very difficult for you to see anything there similarly band to this is band 3 and this is band 4 the reason is that our monitors this support 0 for black and 255 for white so if the maximum gray value is the order of 60 or 70 then you practically do not see anything in the image you practically do not see anything in the image so usually for these images those people who are acquainted with image processing.

They would know that one needs to do some sort of an enhancement so here what I have what we had done was a simple enhancement scheme has been considered so that the images can be seen properly so the enhancement scheme is simple linear stretching for each image you find the minimum gray value and the maximum gray value.

(Refer Slide Time: 03:24)



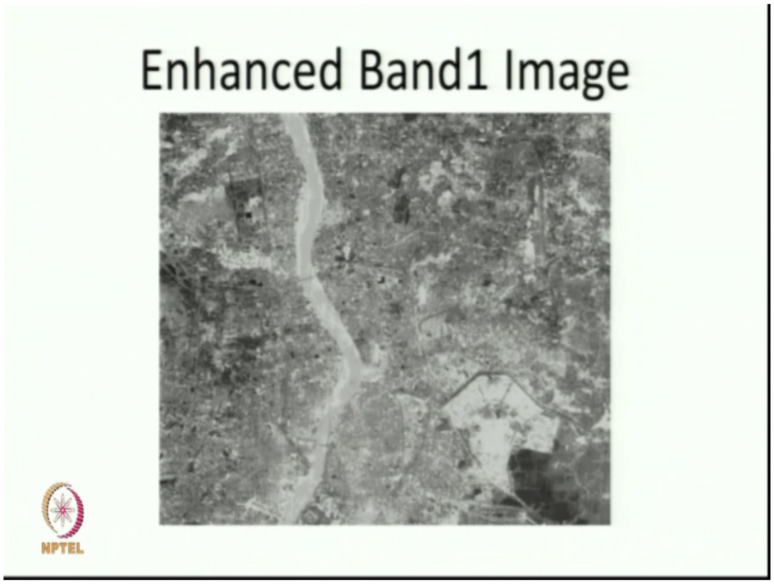
Enhancement

- Linear stretching

$$\frac{x\text{−min}}{max-min} \text{ X } 255$$

And from each pixel value subtract the minimum and you divide it by maximum minus minimum and * 255 so that the minimum gray value in the original image it corresponds to 0the maximum gray value in the original image it corresponds to gray value 255 in the transform domain in the transformed image so since it is a linear function it is called linear stretching since you are stretching the interval from min max to 0 and 255 so you are stretching the interval so it is called linear stretching a simple enhancement operator.

(Refer Slide Time: 04:10)



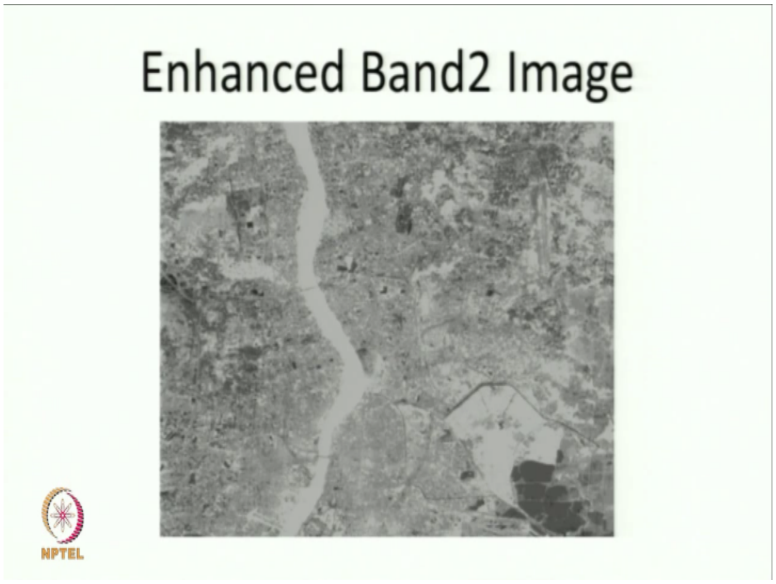
Enhanced Band1 Image

Now after you do the enhancement this is what you are going to get the enhanced image of band one could look like this I will show you some of the characteristics of this thing you see this portion this is basically the river Hooghly our Ganges you see there is a bridge here this bridge is known as valley bridge and then when you come down here you see another bridge this is known as Howrah bridge okay.

This is hardly is the famous Outer bridge and his Bali bridge and it is going down like this and this was the second Hooghly bridge across under construction in those days so and these black things they correspond to pond areas okay this is also this black is also pond area and this black it is also pond area the people who are acquainted with Calcutta actually this is a pond practically in front of ISI when you cross the BT road you will get this pond.

So basically Indian Statistical Institute is situated somewhere here and we are very close to Bali bridge so this is banned one image it corresponds to the blue, blue, band blue color this is the enhanced image for band to that is green please note that the river water and pond water they are different this is the river water and these are this is the pond water this is this blackest pond water this black is also pond water and as I said this is also pond water river water and bland this pond water note that they are different.

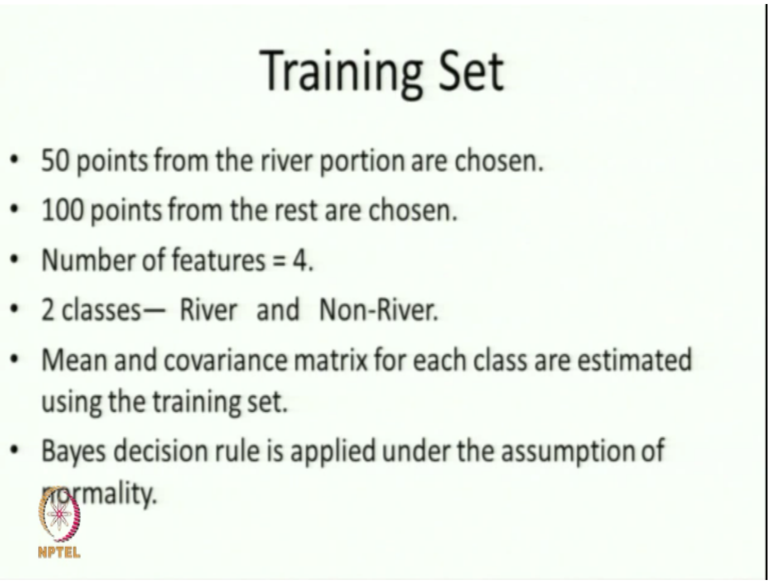
(Refer Slide Time: 06:07)



Enhanced Band2 Image

And this is band to that is green image enhanced one here also as you can see the river water and pond water they are actually different they are giving different gray values and then as you come slowly to this is red.

(Refer Slide Time: 06:22)



Training Set

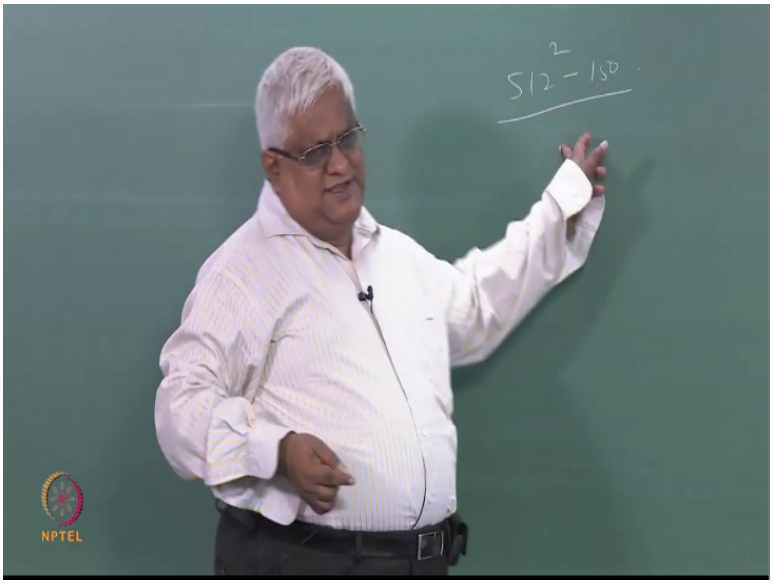
- 50 points from the river portion are chosen.
- 100 points from the rest are chosen.
- Number of features = 4.
- 2 classes— River and Non-River.
- Mean and covariance matrix for each class are estimated using the training set.
- Bayes decision rule is applied under the assumption of normality.

Now let me talk about training sets the what is the problem here the problem is you need to classify each pixel into the land one of the land cover types what is the land cover types under consideration here, here I want to do a simple problem where each pixel is to be put into one of the two classes one class is river water another class is non river non river includes the pond regions as well as you have land and land me includes buildings and you have barren land you have gardens okay and all the other things.

Basically I want to put each pixel to one of the two categories one is river other one is non river this is the basic classification problem under consideration so we have taken a training sample set 50 points from the river portion we have chosen100 points from the rest they are chosen now the question is how they are chosen if you remember my lectures about the how to take training sets and test sets one of the problems that one of the things that had always mentioned there that training set should reflect the overall variation in the data set.

So that you can estimate the parameters properly test set again should indicate the overall variation in the data set so that whatever miss classification rate that we are getting on the test set that should reflect the overall miss classification in the entire population so both the training and test sets they need to reflect the overall variations within that whole population properly so here basically the test set means every point we are going to do the classification so that is basically the test set I mean so the training set is we are going to take in the whole of they are totally 512 square pixels in this 500 pixels we are going to take some small number of pixels some 50 from the river portion and 100 points from the rest this is the training set and the test set is all the rest all the rest.

(Refer Slide Time: 08:56)



I mean it is 512^2 -150 this is the test set size and the training set is100 points from the non River portion 50points in the river portion now there can be first many questions how these numbers 150 are arrive rate that is one question and the second question is once they are arrived at how actually someone can choose these points now the answer to the first one how these numbers 150are arrived at we know that the river portion is small compared to the land port the non river portion.

So the number of points that we are going to take for the training set for the river it will be smaller when you compare it with the number of points that we can take for the non liver portion okay then once it is smaller than this proper I chose1500 some people can even take 25 and 50but probably it should not be less than25 because you would like to get the overall variation in the river class and were all variation in the non river class probably should not be less than25 since I have been dealing with this data set.

So I do not really think it should be less than 25 because you would like to get the overall variation in the river class and of course the overall variation in the non river class so I think one should take at least 25 points in the river and since you would like to get overall variation the non river class also there it is 25 means it should be at least 50, 55, 60 so I have taken 50 and

100 so one can take even different values also instead of 50-100 one can take that as SS 75 and 150 75 and 1 25 or 50 + 125one can choose many other values.

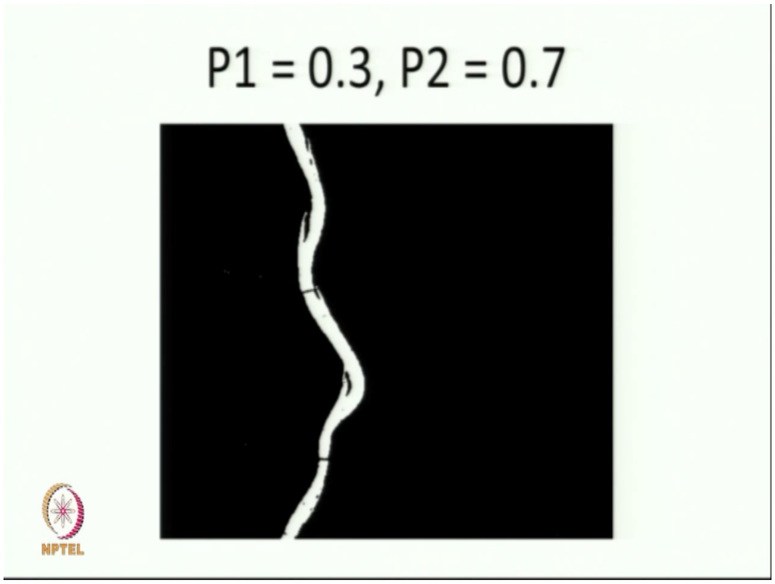
So I have taken 50 and 100 now the next question is which 50 points in the river we have to choose which 50 which hundred point from the land from the non river portion do we need to choose the answer is as I said it should be it should reflect the overall variation now let us look at this portion this is a river portion so if you want to take 50 points those 50 points should be distributed from this place to this place then we are basically reflecting the variations in the class.

Similarly the other 100points for the land you should try to take them as much as possible from all these regions river is slightly easier but the land portion is slightly difficult which particular hundred points you need to choose the other not exactly land portion the non river portion it is slightly difficult in the sense that which exact 100 points are you going to take that is but nevertheless 100 points are chosen now what are the features since for each pixel location you have gray value for a blue band, gray value for green gray value for red and gray value for infrared.

So each pixel location is represented by a four-dimensional pick a feature vector so the number of features is 4 so we have chosen 50 points from the river portion 100 points in the non river portion number of features is for then there are two classes as I said river and non river then after these 50points are chosen then you have 54 dimensional vectors for River 104 dimensional vectors for then on river portion so you calculate mean of these 50 vectors for the river covariance matrix for the river portion similarly you get mean and covariance matrices for the non river portion based on the hundred points.

Now you have the wave mean the mean and covariance matrices note that we do not know anything about the probability distribution so we will assume normal distribution so once we assume normal distribution then we need to know the prior probabilities for applying the base decision rule so for prior probabilities we have taken three cases one case is for the one case is the river portion the prior probability is 0point 3 the non river portion prior probability is 0.7.

(Refer Slide Time: 14:21)



P1 = 0.3, P2 = 0.7

So river portion the prior probability is .3 non river portion the prior probability is 0.7 so and using that we get the base decision rule then you take every pixel get the four-dimensional vector and classify each pixel to either to water either to non river portion or non river portion river portion is represented by the gray value255 in this image the non river portion is represented by the gray values 0 in this image.

So once that is done this is the output as you can see the two bridges the Bali bridge and howdah bridge they came out properly okay the two bridges this is also one of the reasons in my class also I give this example since the river portion is known to all of us one can easily check how good or how bad the classification is you do not need to look at your miss classification rate etcetera one can find it from here directly one can find all these things from here directly you see in the river portion there are many black dots okay.

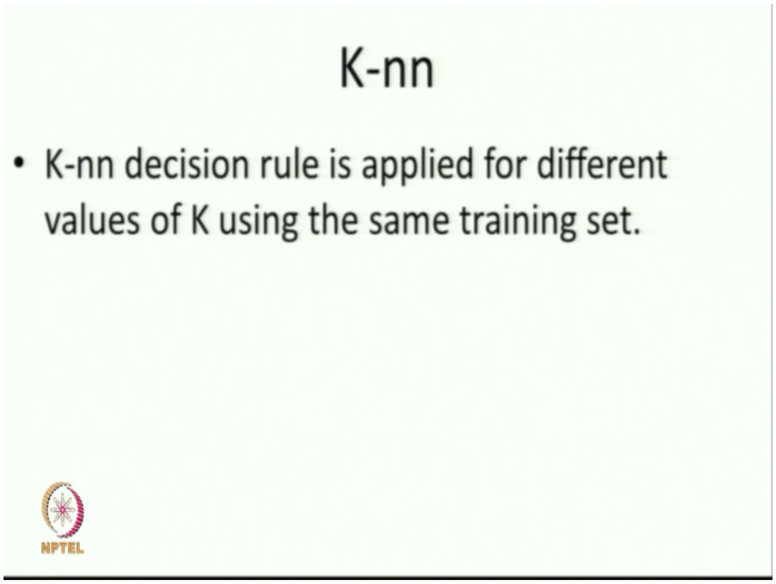
So these points are misclassified but then you see there is a patch here this patch is corresponds to their the river got dried up and this patch again it corresponds to the situation where the river got dried up okay and there are some small, small dots here so these dots they are all misclassified they are all misclassified but note that there are some points here and there are some points here the non river portion they are coming to diverse so there are a few points but then as you can see most of the points most of the pixels they are classified properly out of these many pixels out of these many pixels less than one percent of them are misclassified less than one percent.

So the rate of correct classification is really, really high now before I go further let me make a few more comments then, the first comment is that if you ask me whether similar results will be obtained with other training sets I would say it is not guaranteed I am repeating the same thing again if the training set reflect the overall variation then your quality of result will be good if they do not reflect then the quality of result will not be good it is not guaranteed that the results will be good.

Since the same data have been using for two passed many, many years and every argue this one as an assignment to my students in is I some people they show me results which are worse than this since they have not taken the training sets properly if your training sets are not chosen properly you are not expected to get good results so this one I am telling you repeatedly if the training set is really chosen properly now you see 4.5 and .5 these are is the result and 0.7, .3the results since the training sets is chosen properly the prior probabilities are not making much of a difference whereas if the training set is not chosen properly.

The prior probabilities are going to make lot of difference this I have seen very many examples this I have seen very many examples so this once again reflects that your training sets how to be chosen and there is no they are no compromises here training sets have to be chosen properly.

(Refer Slide Time: 19:02)



K-nn

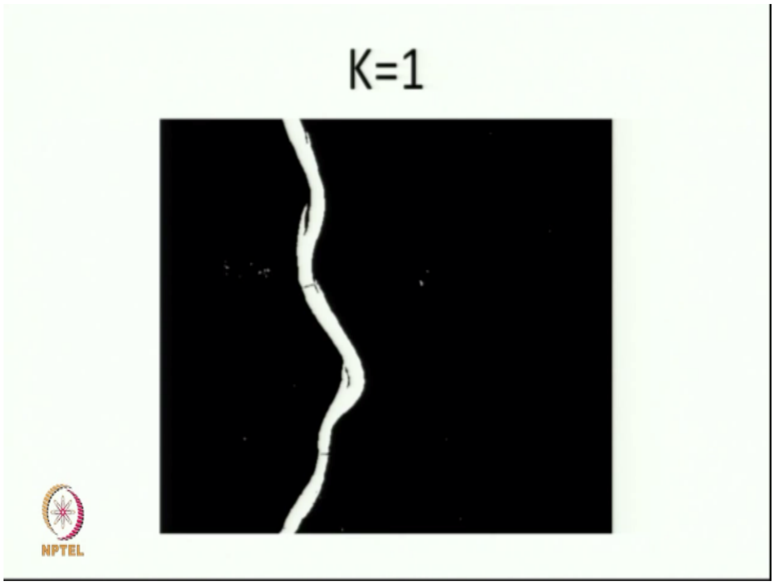
- K-nn decision rule is applied for different values of K using the same training set.

Now the k-nearest neighbor decision rule now you can apply any reasonable decision rule also we have 50 points from the non river portion 100 points in the non river you take them as your training set. So you take one point say the pixel 11so it has a four dimensional vector that pixel is to be classified to one of the two classes so from that pixel you calculate the distance of that pixel with all these 150 pixels what is the meaning of the distance here the distance is Euclidean distance.

And you have four dimensional feature vector so you calculate 150 distances and arrange them in increasing order or non decreasing order and choose the first k distances and you look at the corresponding K points see how many of them are class 1 how many of them are class too whichever value which from whichever class you get the maximum representation you put the point into that class.

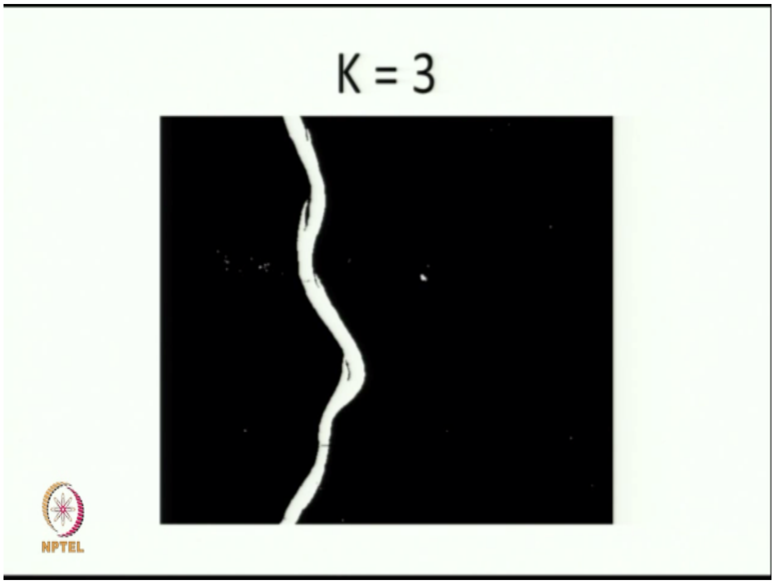
So since here we are only talking about two classes and if I take K as an even number it is possible that you may get from both the classes equal representation so I have to chose you know my value of K to be awed and we are taking the same training sample search this is the result for k is equal to 1.

(Refer Slide Time: 20:25)



This is the result for k is equal to 1now there is one point that I would like to mention here note that here we are not making any distributional assumption what so ever no distribution of the assumption and look at the quality of the result it is fantastic the number of points that are misclassified it is very, very small this shows you the power of this rule k NN rule.

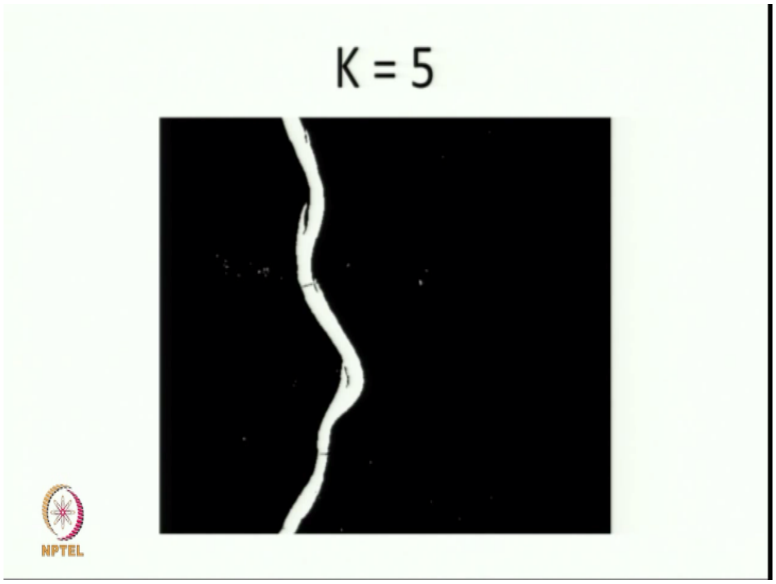
(Refer Slide Time: 21:06)



K = 3

This is for k is equal to 3.

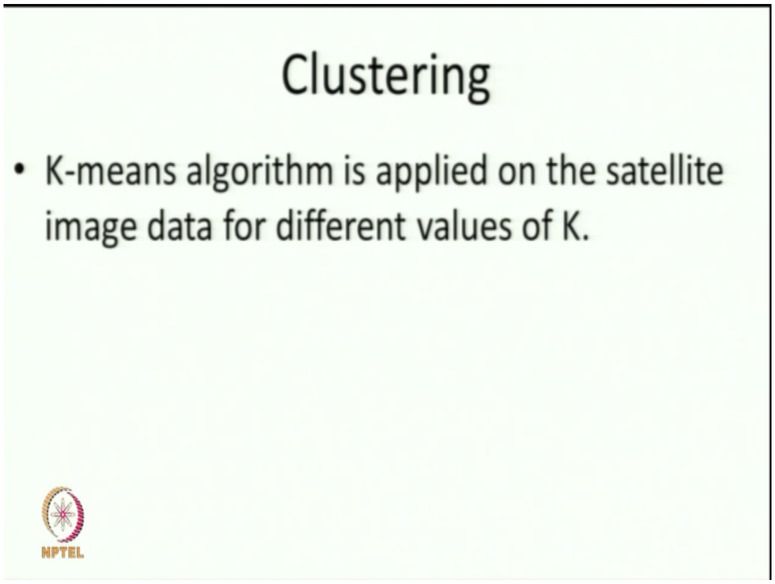
(Refer Slide Time: 21:09)



K = 5

This is for k is equal to 5 so you are able to do the classification properly now there was one question in one of the previous classes about how to choose the value of K how to choose the value of K here I have to I have shown the results on three different values of K there is only very little difference minor difference this again reflects let me tell you the quality of the training sample set if the quality of the training sample set is good for very many different values of K you are going to get similar results you are going to get similar results.

(Refer Slide Time: 22:14)



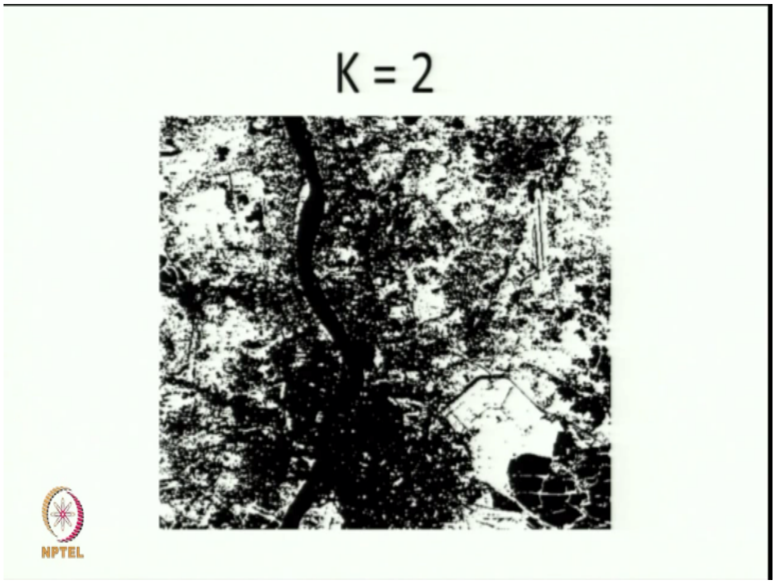
Clustering

- K-means algorithm is applied on the satellite image data for different values of K.

Now this one after we have done the classification then we went into clustering where we have to in many situations we need to tell the number of clusters in advance and we discussed a few clustering schemes one of them is k-means algorithm where k is the number of clusters and for k-means algorithm the value of k is to be given as an input value of k is to be given as an input so on the same data set we did clustering we apply k-means technique and again for k-means algorithm.

There are several versions available we use the 4 g's version that is basically till the whole iteration ends we will not change the cluster will not change anything after the isolation has ended then we have new set of clusters and in the previous one you have old set of clusters if these two clusters are same then we stop it otherwise again we go for next iteration so 4G scheme has been applied and these are the different results that we have got.

(Refer Slide Time: 23:30)

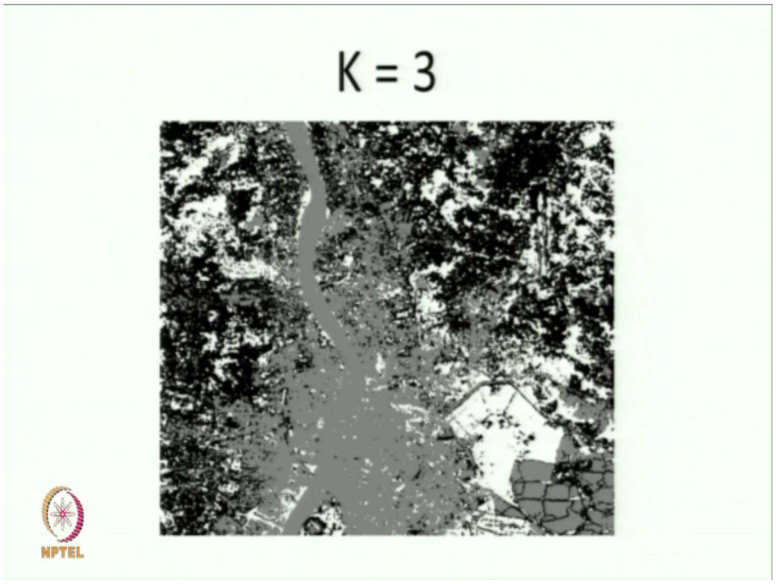


K = 2

Now for k is equal to 2 note that we are doing clustering means we do not know the numbers of groups are number of clusters. We are assuming that the number of clusters is two and then this is the output now how much this output reflects the reality that is the basic question if an output should reflect the reality then we need to get the value of K properly if an output should reflect the reality then we need to get the value of K properly and we do not know the value of K we do not know the value of K so we have tried it with current values of k for k is equal to 2 this is the result.

This does not actually say much about the grouping I mean we are not in a position to name the different clusters.

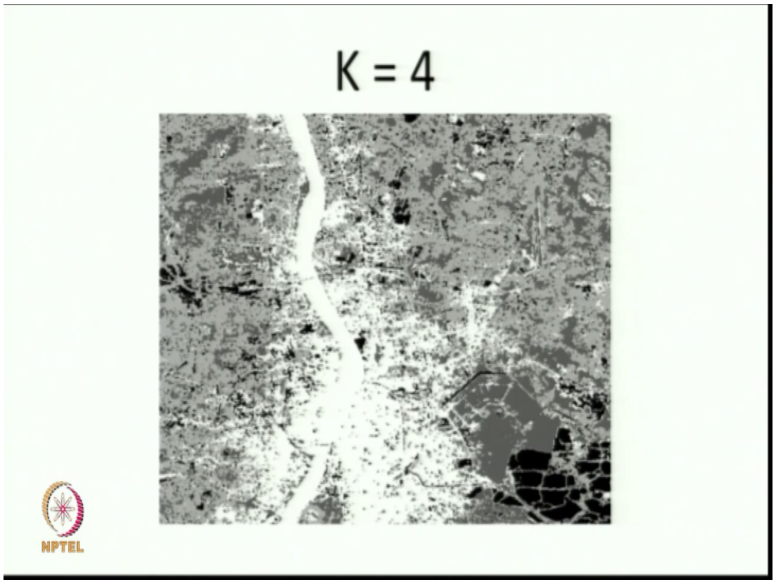
(Refer Slide Time: 24:26)



K = 3

This is for k is equal to 3 this again we are not in a position to name.

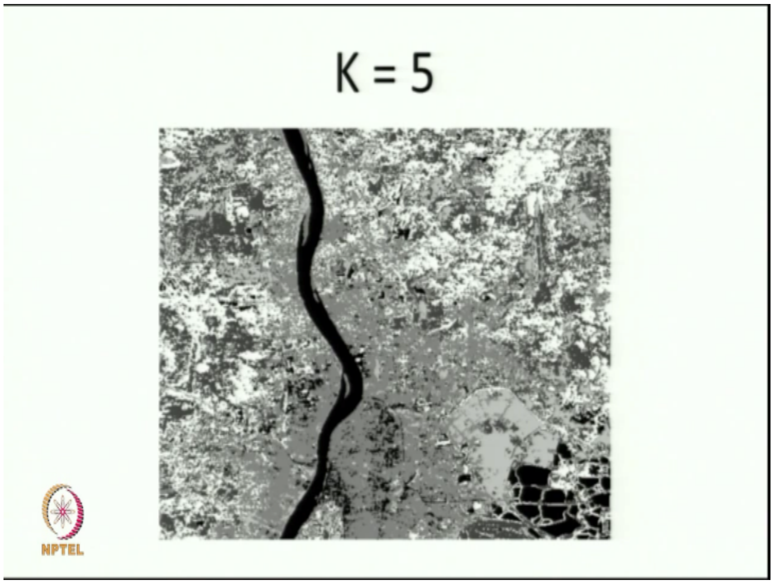
(Refer Slide Time: 24:32)



K = 4

This is for k is equal to 4 this is slightly better but then let me just discuss this.

(Refer Slide Time: 24:44)

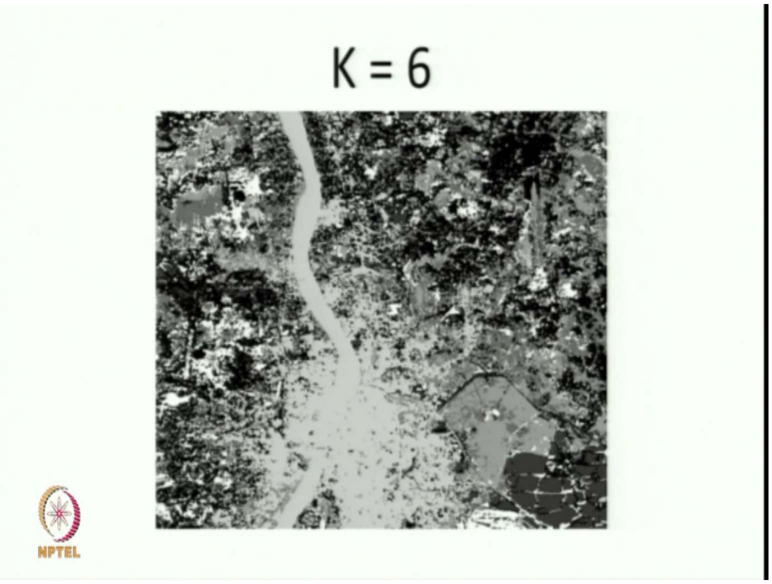


K = 5

This is for k is equal to 5 look at this, this is the river water and this is the pond water and wherever this black color is there wherever this back black color is that most of it is water well I do agree that this portion this is not water this portion this is not water but these portions they are indeed water this portion is water so here these portions they are not water body this is basically the main city area of Calcutta this portion is a sham bizarre.

And then you come down to Schaumburg earth this is how a bridge so there is all somewhere here okay and this portion it is the modern portion this is slightly better result.

(Refer Slide Time: 25:56)



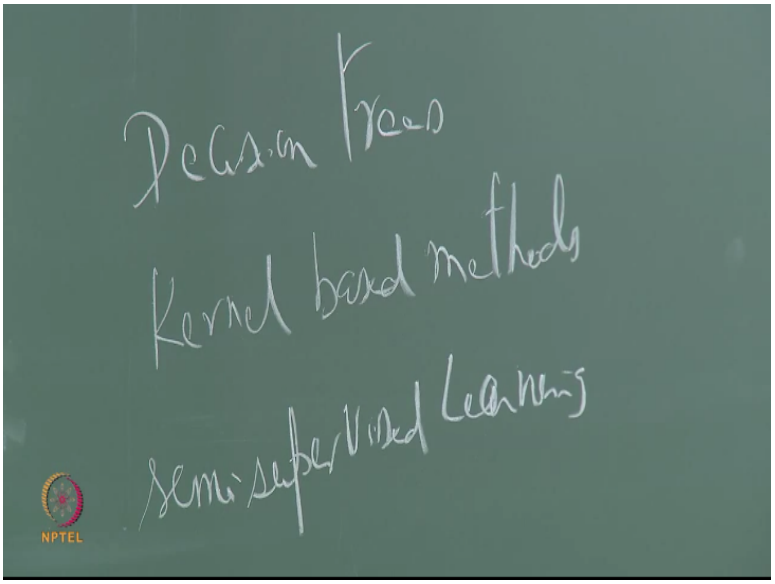
K = 6

This is for k is equal to 6 here river and river and City portion they are mixed up but except that this is the pond water pond water this color you see at many other places also but if you ask me ultimately my conclusion is that they are giving you clustering this clustering is sometimes some places some clusters are reflecting the actual reality but I cannot make a uniform statement about any particular value of K.

I cannot make a uniform statement about any particular value of K that this value of K is better than the other I mean in all respects that I cannot make but these are the results that we have obtained well this is a basic and inherent problem in clustering they will provide groupings in the dataset whether those groupings reflect the actual reality our actual names that we are going to give to the clusters that is not necessarily true they reflect the groupings within the data set then they may not correspond to the names.

That we would like to give so we have actually come to the end of the lecture series on pattern recognition we try to cover several topics in classification clustering and feature selection in pattern recognition but there are still many more topics that are left uncovered one of them is syntactic pattern recognition that is one and among the usual techniques some texts that have not been covered are you have decision trees.

(Refer Slide Time: 28:03)



Decision Trees

Kernel based methods

semi-supervised Learning

You have decision trees and kernel based methods kernel based methods in machine learning and you have other machine learning schemes like you have semi-supervised learning reinforcement learning and you have poison forests okay and there are like that several other schemes are available transfer learning I already wrote semi-supervised learning reinforcement learning and there are many more classification schemes there are many neuron classification schemes available which have not been covered.

There are many classification schemes available using some other soft computing techniques like rough sets and genetic algorithms and in fact there are many, many other schemes available if go on mentioning the names they you will get in fact many more names please take it as a lecture series for beginners please take it as a lecture series for beginners since we have not we have covered only some portions and they are still many other portions to be left covered and whatever you teach here there will be always some portions which will be left out after all the time is finite.

But the number of things that you have to do the syllabus is quite a lot please take these lectures as lectures for beginners only and many advanced topics have not been covered people who are interested in working on those advanced topics they should read the corresponding literature and they need to always compare whatever they read our whatever techniques that they are going to develop or use with they need to always do the comparison properly among the techniques.

That is one thing that I request all the users who are working or who did like to work on pattern recognition and related topics they should do their comparisons properly doing the comparisons properly means you should take that if you are taking the one training set you should use the train the same training set for all the class fear classification schemes you should not change the training sets similarly if you are using certain parameters for something then you should use similar parameters for the other schemes also from scheme to scheme there are suppose you would like to do reduction of features you would like to do reduce the number of figures from 100 to 10 okay.

You would like to reduce number of witches from 100 to 10 you have developed a scheme now with some other schemes from 100 to 10 probably may not be positive maybe it becomes 11or it becomes 9 but it should not become20 or it should not become 5 so when you are making this comparison be it with feature selection or classification are with any other things what you should do a proper comparison that is really, really necessary and nowadays for I mean whatever techniques that you are going to develop if possible.

If you develop the corresponding theory for the techniques then the techniques will last longer since there is theory one will always know the limitations of the method properly if you know the limitations of the method then you would actually know where to use it and where not to use it so if you can develop the theory for your methods properly that will always be good and in that sense the method also will last longer but if you are not in a position to develop the theory probably for a few years the method may be existent.

And after that people will go for the new methods that is why you see in the latest literature for many journals people then, the associate editors and editors they are actually asking for papers with theory the main reason is that with theory you will know the limitations properly having said this I all I mean it, it can be seen that developing theory is not an easy task having said this it is not an easy task to develop the theory.

But for the good of the subject if theory is available it will always be better so for all the users of pattern recognition and for all the persons who like to work on pattern recognition from my side and my colleague professor session to the outside thank you and hope that these lectures will help you and if you have any questions you people can always write to NPTEL and from there we will get those questions and we will try to reply to your queries thank you very much.

Online Video Editing /Post Production

K.R.Mahendra Babu
Soju Francis
S.Pradeepa
S.Subash

Camera

Selvam
Robert Joseph
Karthikeyan
Ram Kumar
Ramganesh
Sathiaraj

Studio Assistants

Krishankumar
Linuselvan
Saranraj

Animations

Anushree Santhosh
Pradeep Valan .S.L

NPTEL Web &Faculty Assistance Team

Allen Jacob Dinesh
Bharathi Balaji
Deepa Venkatraman
Dianis Bertin
Gayathri
Gurumoorthi
Jason Prasad
Jayanthi
Kamala Ramakrishnan
Lakshmi Priya
Malarvizhi
Manikandasivam
Mohana Sundari
Muthu Kumaran
Naveen Kumar
Palani
Salomi
Senthil
Sridharan

Suriyakumari

Administrative Assistant

Janakiraman.K.S

Video Producers

K.R. Ravindranath
Kannan Krishnamurty