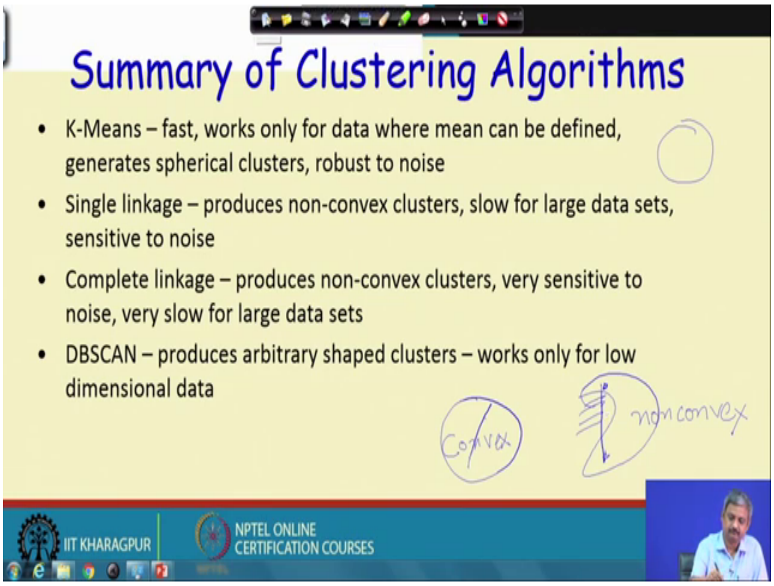
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 36
Clustering – V

Let us compare few clustering algorithms we studied.

(Refer Slide Time: 00:25)



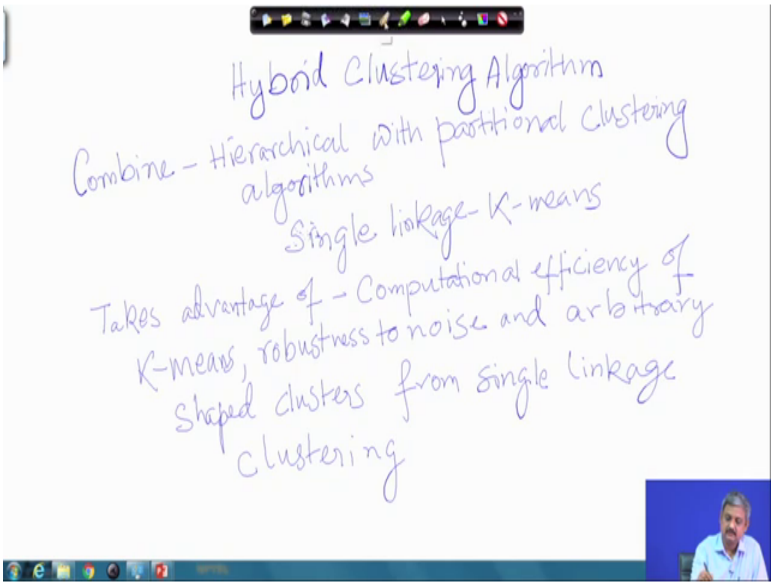
So, we studied three algorithms; the K – Means, the single linkage a hierarchical and the DBSCAN, density based. So, the advantage of K – Means is that is very fast, it is robust to noise, but it works only when the mean can be defined that is numerical attribute it does not work for nominal or ordinal attribute. So, it is difficult to define a mean. Also the disadvantage is it general spherical cluster.

So, if you the natural clustering is non convex, non spherical, by the way this is a spherical cluster would look like this. A convex cluster is anything which looks like this. So, this is convex and this is non convex. So, how do you know what is convex, what is non convex? You take two points inside a cluster draw a line the entire line will lie inside the cluster whereas, in non convex if you take two two points join them by line there will be some points on the line outside the cluster.

Ok. The single linkage provides non convex cluster if the natural cluster in it. It can also provide spherical cluster, but it is sensitive to noise and slow for large data whereas, the complete linkage is also non convex, but it is also sensitive to noise. Actually it is much slower than the single linkage this even more slower than single linkage, but it produces very good quality, elongated clusters if required.

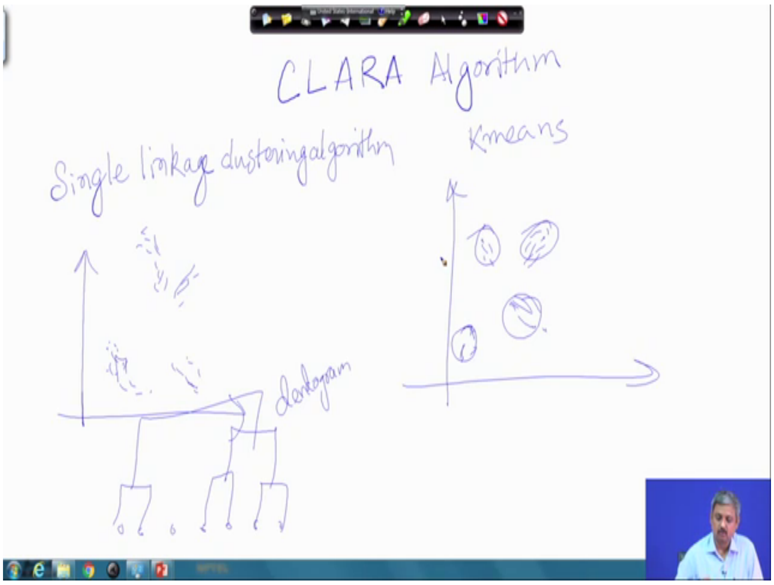
And, DBSCAN it provides density based clusters, you can get any shape, but because of the K nearest neighbour density estimation it will only work when the data dimension is low, less than 4. So, usually spatio temporal data there it works. There are many modifications to this algorithm for example, one modification to the K – means algorithm is the K-centroid or K-medoid, where other K-median also where instead of the mean being updated you take the centroid or the median getting updated.

(Refer Slide Time: 03:14)



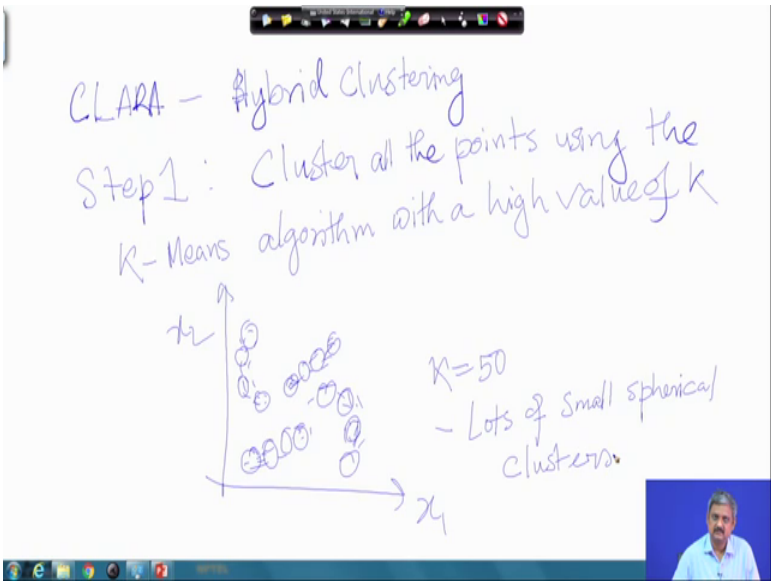
Ok. So, what people have done is that they algorithms say and clustering ok.

(Refer Slide Time: 05:44)



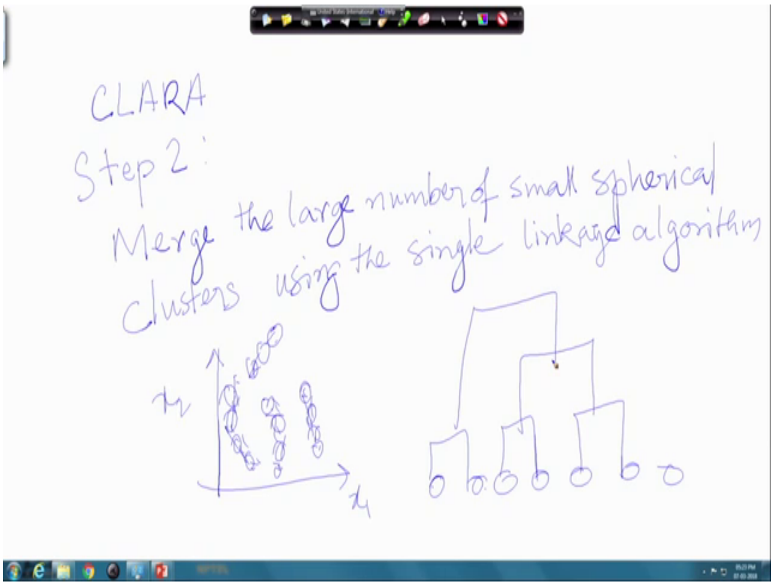
So, this algorithm is called the ok. So, if you remember what the does is it merges this point and creates a dendogram, whereas, K – Means what it does is that it just forms a partition updating the mean, ok.

(Refer Slide Time: 07:20)



So, CLARA combines this two. How it combines? Choose a high value of K and run the K-means. So, these are really non convex clusters. So, you choose say K equal to 50 even the natural cluster is only 4. So, you choose K equal to 50 let lots of small small cluster, clusters clear. So, run the high value of K with K – Means get small small circles to cover up the data oh, sorry merge the.

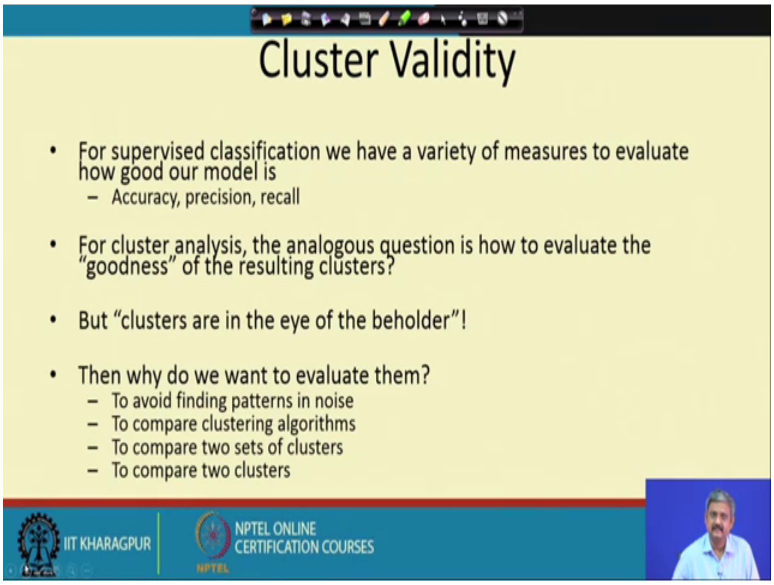
(Refer Slide Time: 09:33)



So, the points like this the first step you have what I do is I will merge them by, so, in single linkage interrupts single points swing leaps these small clusters will be leaps and maybe distance between them is distance between their centre of this point, the spheres merge.

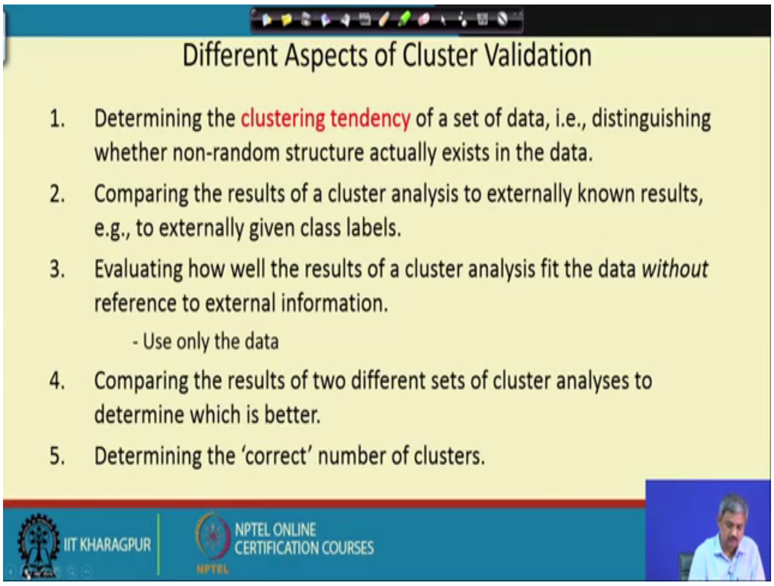
So, here this since number of spheres are much less than the number of points, single linkage takes less time and also you get non convex cluster, ok. So, that is the idea. So, there are many other algorithms like that this is just an example. So, now, let us come to our next topic.

(Refer Slide Time: 11:34)



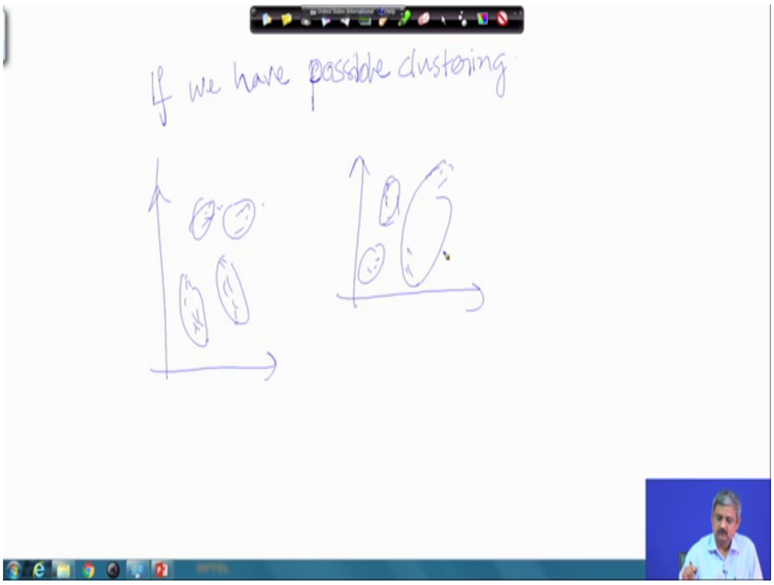
Evaluating clustering algorithms, how good are they? You have so much variety, which one do you choose? ok. So, for classification we have seen accuracy, precision, recall what do we do here, but the important thing is clustering depends on what is good cluster depends on application, but still we need to compare them because of this these this four reasons, ok.

(Refer Slide Time: 12:15)



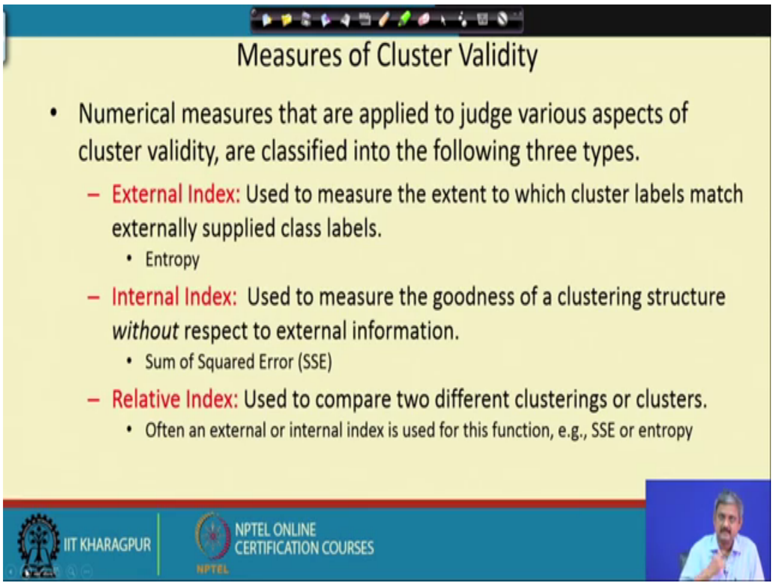
So let us see you have to these are the regions, tendency. Can you compare with existing class levels?

(Refer Slide Time: 12:54)



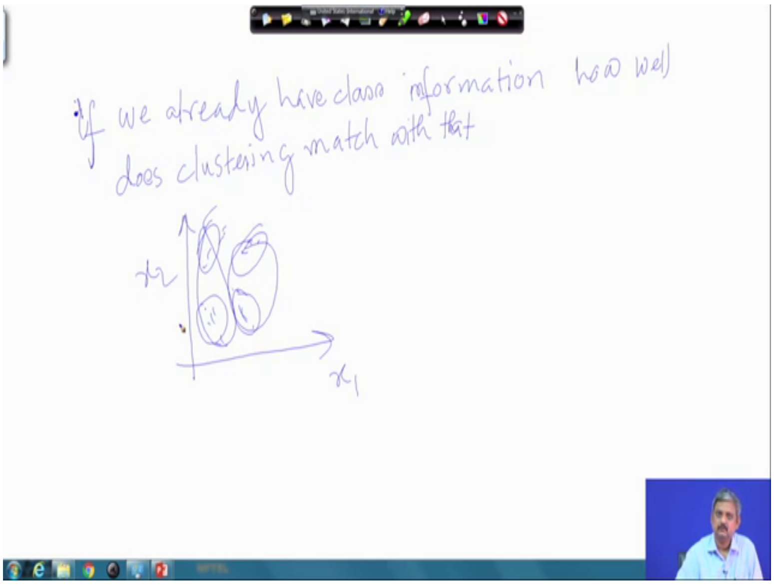
Using only the data, how well can you evaluate? How does two clustering compare? So, this is one, this is another, ok. How good that, how does they compare?

(Refer Slide Time: 13:42)



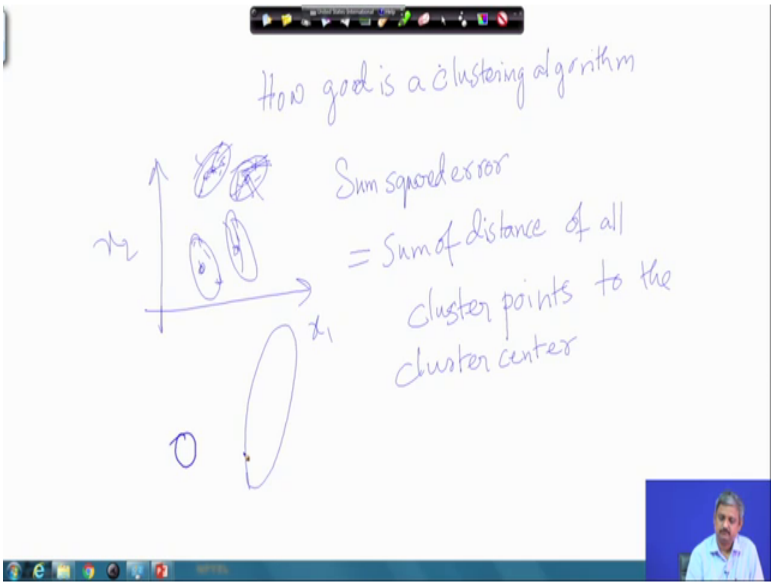
There are many measures called cluster indices index, external, internal, relative. External means there is already some ground truth some class level, how can we compare against them.

(Refer Slide Time: 14:10)



Meaning, suppose class level is this already known; given. How well if you cluster like this it matches with that? That is external.

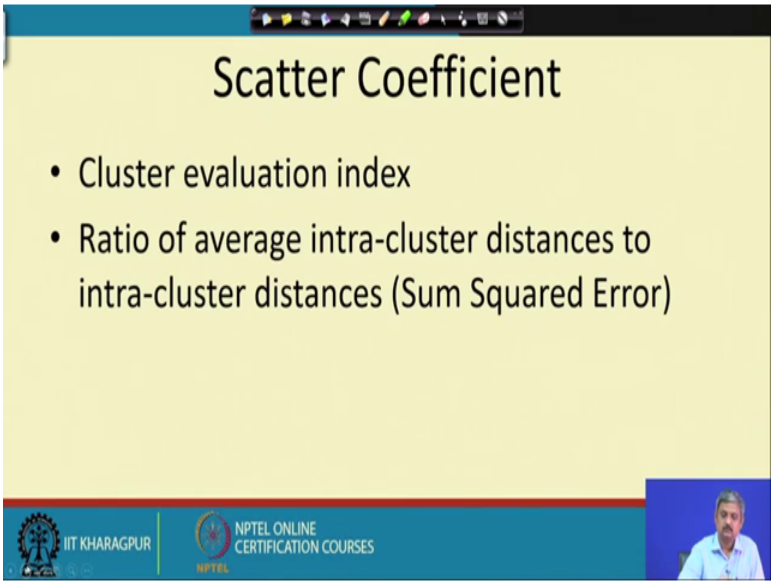
(Refer Slide Time: 15:18)



Internal without knowing class level, how much can if I may say sum square error; so, in K – Means if some of this part these distances, ok.

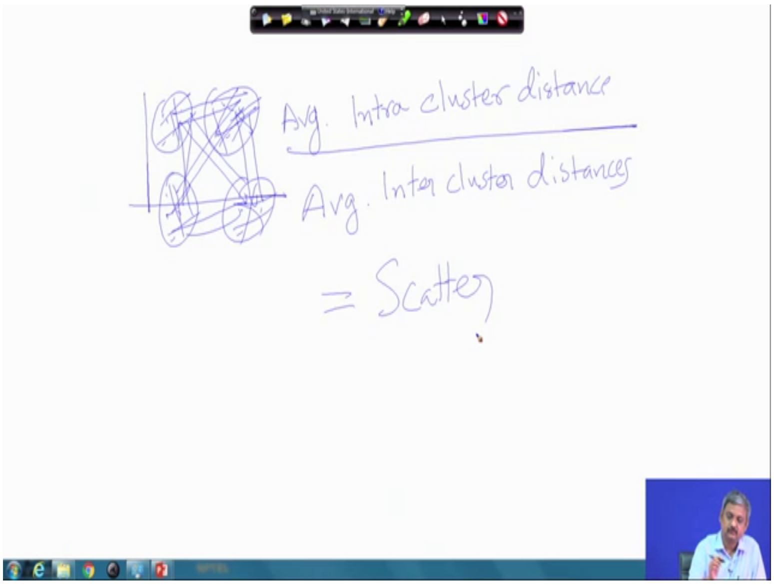
So if the clustering is dense it will be low, it will be good. It error will be low, it is high, it is not good and that is a relative performance.

(Refer Slide Time: 16:43)



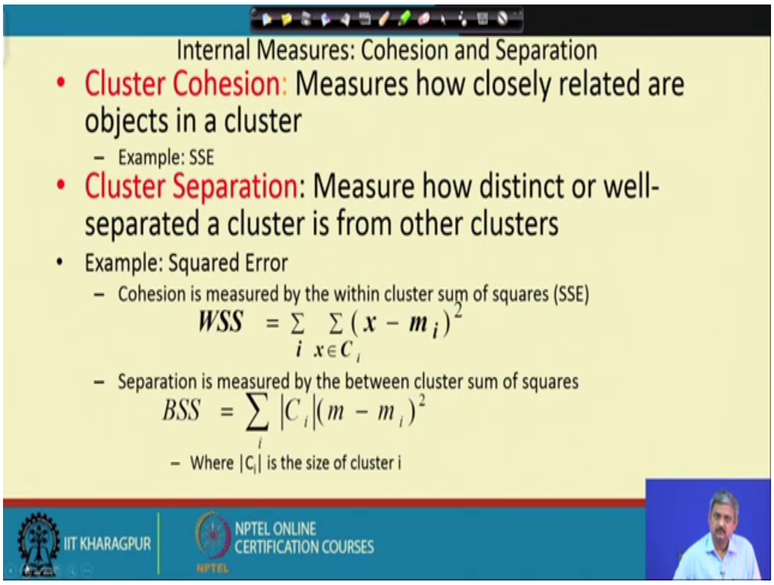
So, the most common squared sum error is the scatter coefficient, which we discussed.

(Refer Slide Time: 17:07)



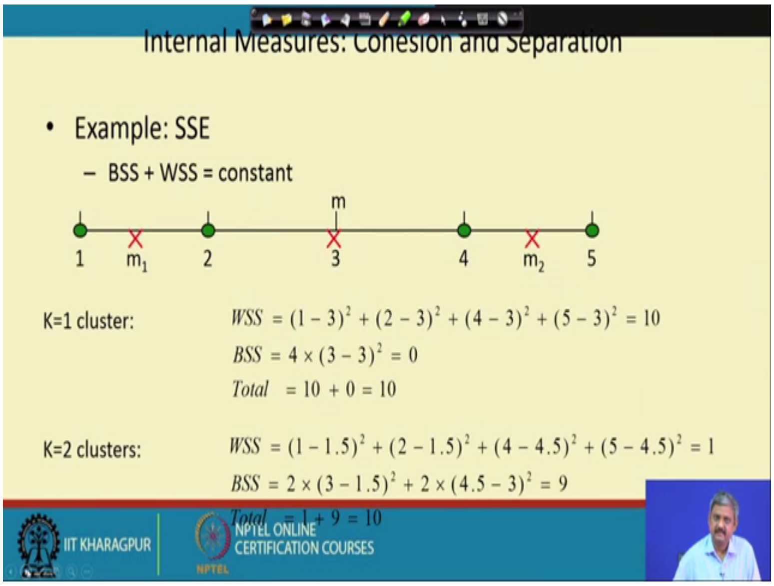
It is nothing, but the ratio of so, it is the; take any two points lying in the same cluster, this distances average of that, all these distances. Is the scatter, so, that is what we have defined.

(Refer Slide Time: 18:18)



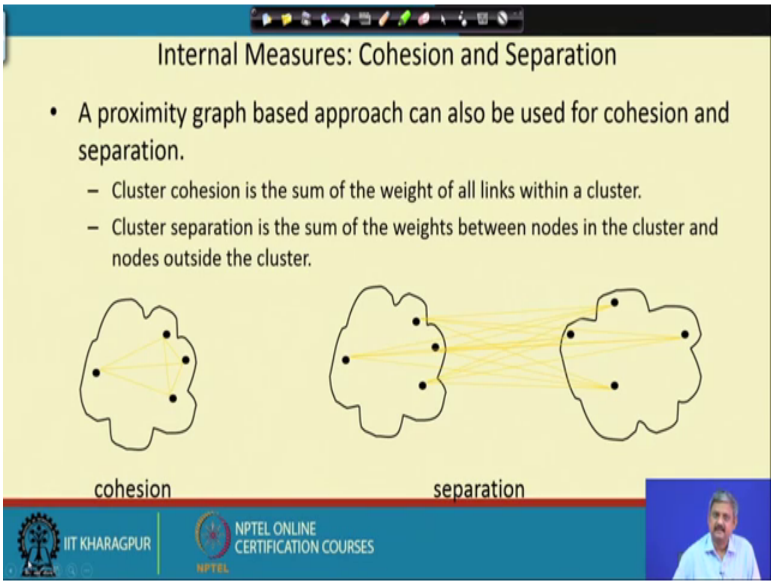
So, these two terms are as I have told you these distances, these are called cohesion and separation and scatter is ratio of cohesion to separation.

(Refer Slide Time: 18:54)



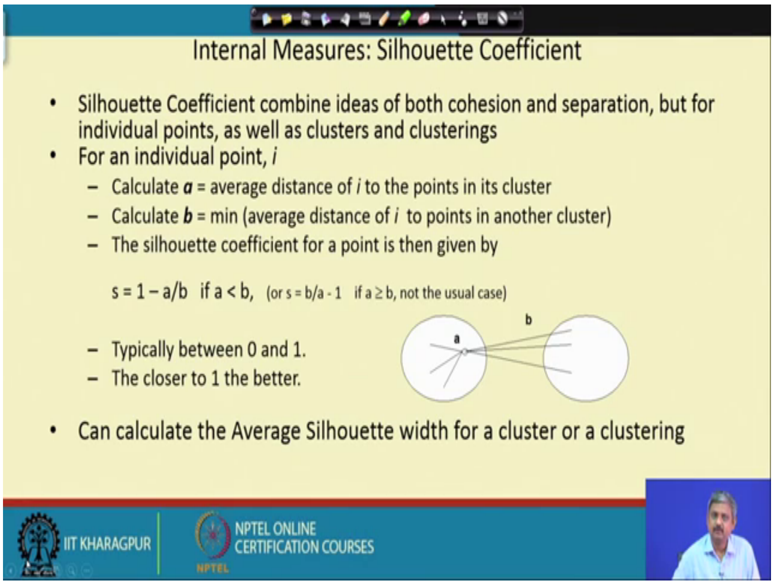
So, here is an example. So, this is within sum of error between sum of error. Here is an example of how we are getting these values, ok. So you, so, if you take one single cluster this is the mean this is the value, and if we take two clusters, red and green this is the value. So, you can use it to compare, ok.

(Refer Slide Time: 19:35)



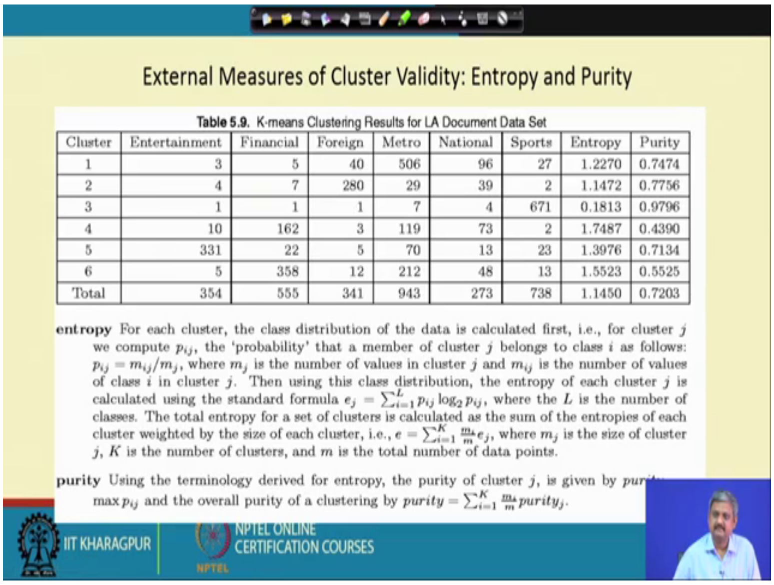
So, as I have mentioned.

(Refer Slide Time: 19:42)



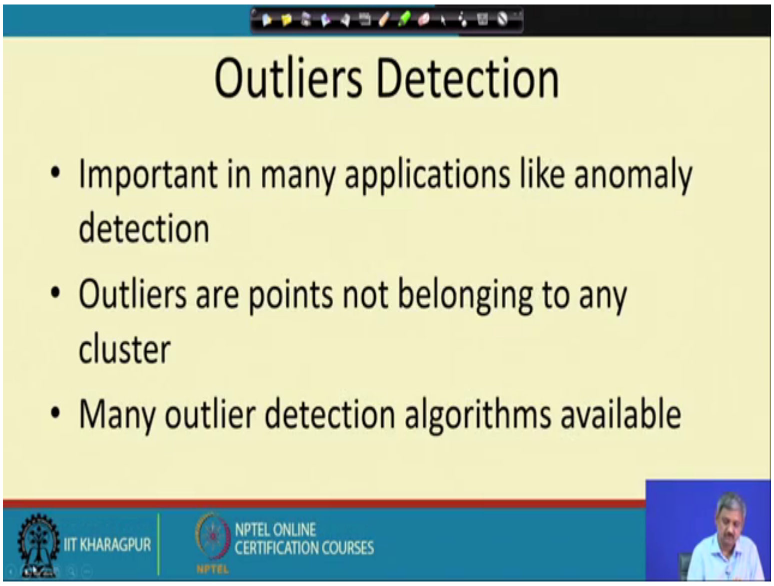
There is another important coefficient called the silhouette coefficient, ok. The silhouette coefficient, which is combination of these two so, you take this sorry. Similarly, you take this and if a is less than b silhouette is this, if b is less than s silhouette is this it is between 0 and 1; closer it is to 1, the better and you then calculate the average silhouette for a all the clusters. So, you can calculate the silhouette of this ok, how to get that, all right. So, this can be calculated.

(Refer Slide Time: 21:19)



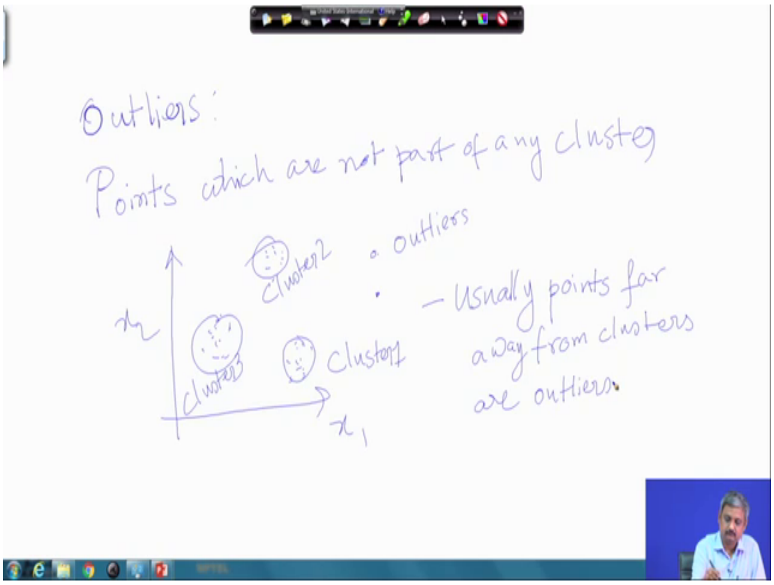
There are some external measures also, which are entropy another which are defined here. This is an example for some data set entropy and purity, like that is (Refer Time: 21:33) we have discussed earlier you can consider that also.

(Refer Slide Time: 21:36)



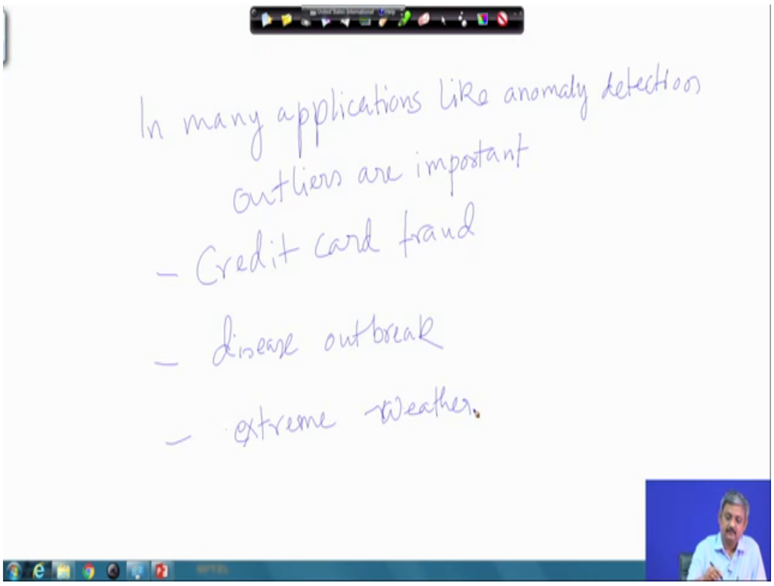
Another important related thing to outlier is into clustering his outlier detection this kind of complement of cluster, this is not a cluster, ok.

(Refer Slide Time: 22:12)



So, in many applications, actually outlier detection is more important than anything. So, ok, these are outliers.

(Refer Slide Time: 23:34)



Important things like this ok. So, that is what is outliers.

So, with this I complete my discussion on the clustering and outlier detection, an important class of unsupervised algorithm. In the next lectures we will look at other type of data mining tasks.

Thank you.