NPTEL

NPTEL ONLINE CERTIFICATION COURSE

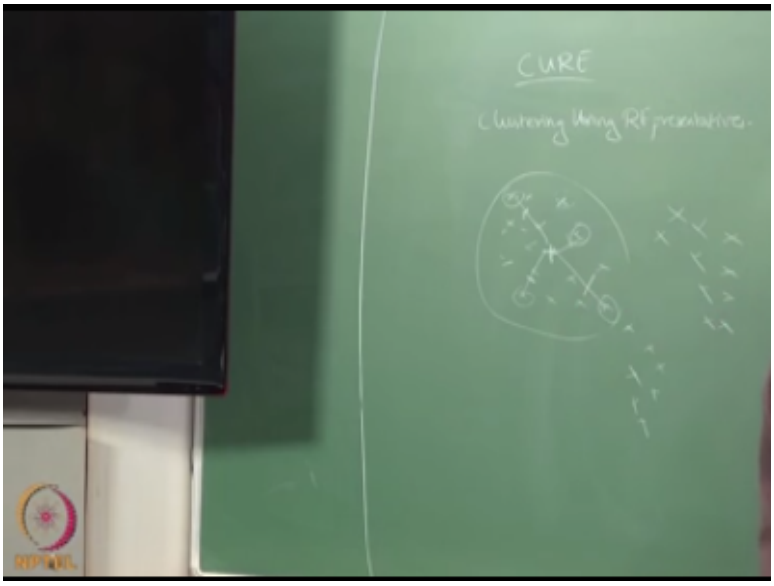
Introduction to Machine Learning

Lecture-74
The CURE Algorithm

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Cured sorry it is its abbreviation yeah the clustering using representative points I got that now is that no represent just use the first two letters of representative representatives okay, cheer up sons bad right cure sounds is nice and this is clustering using representatives right.

(Refer Slide Time: 00:57)



Okay fine so with CURE again was touted as a an algorithm for handling lately large datasets that CURE again this is an algorithm for handling large data sets, so the way you do this here unlike Burch where you actually go through all the data points so CURE what you do is you sample a large fraction of the data okay, as large a fraction as would fit in your memory comfortably sample a large fraction of the data and then you do some kind of clustering on that right.

So what you do in the clustering then you after you've done your clustering right so you start off with some initial clustering of the data right, then what you do is let us say I have something like this right some kind of weird shaped data like this side once I have done some kind of clustering so I will take the center okay, I will take the data point farthest away from the center right so this is let us say this is a cluster I take the data point farthest away from the center let us say that is it okay next I take the data point that is farthest away from this data point.

In that then I take a data point that is farthest away from roughly do not worry I am not measuring it with a scale or anything right I will take the data point that is far the farthest away from both of them put together right so I compute the distance from both and then add it up and then, let us say that one and then I figure out one more if I need it one more right I will take that one right so now I will take these as may represent that one and take these as my representative points for the cluster.

Right so a data point gets assigned to that cluster which has a closest representative point basically I am looking at the boundary points, right I am looking at the representatives or things that delineate the boundary of my cluster right I do not take too many points right I take some number of Representatives lately like 3 or 4 representative points at all or say ten representative points so that they do not completely trace out the boundary but they give me some kind of idea of what the boundaries okay.

So this is why representatives I reassign the clusters I reassign the data points right so I have some clusters, now I find the representative points I find the representative points for this cluster I find it for this cluster now I reassign the data points to clusters so a point goes to the cluster which valve whose representative point is closest earlier we used to reassign them to the close a Centroid now will be assigned them to the closest representative point, now for each cluster I will have 4 points right for each cluster I will have 4 points.

So what I will do now is I will forget the clustered memberships right this is just like how k means works right so k mean what will I do I will reassign a data point to the closest centroid now I will reassign the data point to the cluster that has the closest representative point initiate whatever the whole thing is done out of a set of samples, have a set of samples I drew from the

data for points for each cluster Oh yes, among the sample post but this one thing I forgot about the representative points1 point I forgot about the representative points apologize.

So once you identify these guys right so you do not take them as your representative points because they become too susceptible to outliers okay, you do what is called shrinking right you move them some fraction α towards the centroids, let us I find the weight the points at the boundaries right and then I shrink them a little bit towards the center right so that I do not get too influenced by the outliers, and the shrinking is done by a fraction of the distance of the centroid so that the farther away for you are from the centroid.

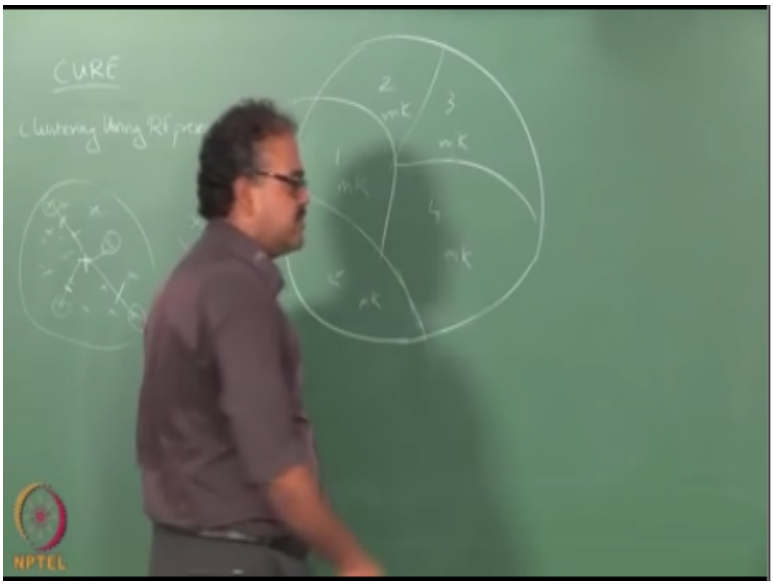
So the greater the shrinkage okay, so this hope you find the representative points and once you find the representative points you go ahead and keep doing this on the same sample right. So the idea of taking the sample is the sample is small enough to fit in your memory right, so that I do not have to go back to the disc to read the data for the second iteration they still a couple of stages to go this is a first stage of cure at all sample I cluster using this representative points okay, so remember that the sample.

I hope is representative of the whole data set but it might not be right I have taken some as large a sample as if can from the data right into my memory and then I do this clustering around representative points right is it the clustering the representation based clustering is clear right, just like k-means instead of assigning it to the data point with the closest centroid I send it to the data point with the closest represent at it is a parameter yeah cm some m you choose like in this case we chose it to be four right, yeah was it as I said you start off with an arbitrary clustering you start off with some arbitrary clustering okay and then you start pick representatives find the centroid of that clusters that you have right and then go to the corners shrink them right do not forget the shrinking step.

Shrink them and then you get a representatives right, so just like the centroids are not real data points the representatives also need not be real data points okay because once you shrink them they are no longer so when you find them they are real data points then you shrink them then they are gone right they are no longer real data points okay, now you get m such sorry I will be using all the data points of the point but only the sample so we did not bring the clustering I am not worried about anything else.

That I have not sampled right now what I do is I keep doing this until I have converged to some clustering right once you have converged to some clustering, okay what I will do is I will remember those representative point suppose I have k clusters I will have M times k points it will remember those k representative points I will forget everything else right, now I will pick another sample from the data I will repeat this in fact the recommended way to do this is to partition the data initially randomly into some large number of bins like say k bins at partition the data randomly into k bins and then do run cure in parallel on each one of those cabins.

(Refer Slide Time: 08:47)



So if you have a lot of machines so what you can do is you can take your entire data set, okay this is not meant to be a geometric representation of the data right, so now when I say I am going to divide it in two parts I am not saying take data that belong to one part of the input space and assign it is I do not know this right this is randomly split the data right and say that okay here there are 5 bins again in each bin I will run cure independently right and what we will end up with some mk points for each of these bins.

Now what will I do is I will throw in all the mk points into a single clustering problem and one CURE on those mk what about 5, 5 times mk points again, again you will get a set of representative points right for each cluster will get a set of representative points then I will go back and I say each data point to the cluster which has the closest representative point right and

on each part I run CURE right I will end up with some clusters right and for each clusters I will have a set of m representative points.

Let us say I end up with k clusters in all of them I will a priority define k right I will end up with k clusters in all of them right and so there will be some mk points from here where m is a parameter that I have chosen already which is the number of representative points let us say for now what is happening is I have a small fraction right for each cluster I am going to have 4 points is a small number with my number of data points will be very large the number of clusters will be small right I will have 4 times.

The number of clusters a small number right I have now 5 such divisions so I will have 5 times m times k points case number of justice k clusters just like a means you define k theory, so but the point here is at every point at no point am I looking at a larger clustering problem right so I will split the data right so I am looking at some fraction of the Ray one-fifth of the data or one-tenth of the data and that is the largest data set size I am going to look at this is why CURE is a algorithm for handling very large data.

So I can take a very, very large data I can split it up 10 ways second spirit of 15 ways 20 ways and I will do each one of these clustering's so if have if I am looking for 100 clusters and for representative points, so that is basically 400 points returned from each one of these clustering's right and then what I do is 55 of these right so I basically end up with 2000 points right so I have100 clusters for representative points per cluster right and 5 such problems I have solved so end up with 2000 points.

Very, very small number right so I go and run cure on that again right again I will get 100 clusters on this 2000 points right, bad idea 100 classes on 2000 points is a bad idea basically end up with 20 points per cluster, so what you should typically be doing is whatever final clustering you want to end up with this should be much larger than that, this K that you use here suppose I want to end up with 30 clusters then I can use 100 clusters suppose I want to end up with 100 clusters I better end up using a larger number 200 clusters or something okay.

So that I love enough data points to cluster finally and then I do the clustering suppose I end up with 30 clusters each will have 4 representative points I will pick those 4 representative points I will do another scan through the entire data, set right how many scans I have done through the

entire data set so far one so far I have done only one when did I do the scan to do the partitioning right, so I did the scan wants to do the partitioning the second time I do the scan I will have m representative points from 30 clusters.

And I will go back go and assign the data point to the cluster that is the closest representative point right, so at every point I do not solve the problem that is larger than what I put into this one partition this allows me to do these things very rapidly and if I have multiple processors that can run right I can do this in parallel this stage can be done in parallel and whatever cluster representatives it returned and then I do in a second round of thrusting, so CURE is if you do if you do the implementation correctly.

It is rather fast another nice thing about CURE is that because I am doing this clustering around representative points I am not really limited to looking at convex clusters like I have with k-means rights okay means if I have two centroids right so basically this will be the separating FF three or 4 centers I basically be looking at some kind of convex shape around the centroid right, but if we have since you have multiple representative points I can actually have non convex shapes also disadvantage of using non convex shapes q this is overhead right so if the data is small you really do not want to get into the CURE kind of a setup right because the overhead is large.

And you have to maintain so many representative points right so every time finding a representative points involves you getting the centroid and doing this computation to go to the edges and then shrinking them right so it is additional overhead that k-means you just find the centroids and you just move on here you have to do an additional computation, so that is the overhead okay good, so any other any questions on this I said a question is no it usually defeats the purpose I am assigning it to the data point that has the closest representative point right so the centroid is will be anyway surrounded.

By the representative points I am taking so it will anyway end up going to the same there might be a small change here and there but people typically do not include the represent the centroid, in the representation this is that we have to now find a set of two parameters of prairie yes but you get some advantages was it right it runs on large data sets which is stop and think about it you could do and claim it run some plastic clusters right, so the and the second thing is you can get non convex clusters.

So if I have fun funky shapes funny shape clusters then you get those in the sense representative points are not real data points they are fictional points like centroids so when it do the reassignment of points to clusters, representative points never get reassign I mean they are not points at all remember I told you go to the edge you find real points and then you shrink it by a factor α towards the center, so just another parameter we have to select so three parameters now it is drink it by some factor towards the center so therefore they are not real points just like the centroid does not really get reassigned neither do the representative points.

So whole point of doing the shrinkage was the reverse the total layers right, now you could you could think of using the mid idea more computation without replacement picture yeah without replacement that is why I showed the partitioning typically sample without replacement and CURE you could I am not saying you cannot I mean see the point is for every, every variant that you propose you have to either convince me empirically that your variant is better or convince me theoretically that your variant is better.

No that is this let is go feat will generate papers right and let me stop and think what is it that it buys you know is there some kind of statistical parameter that will become better by sampling with replacement is a sampling without replacement for example will you get more stable cluster estimates if you sample with replacement, so regardless of what how many ever how do I speak the samples right so if you say that I will get the same cluster centroids maybe then that is a valid thing to do right if you cannot show anything like that.

Then not clear right so the thing with which has more overhead sampling with replacement or sampling without replacement sampling without replacement has more overhead really with replacement has less overhead if you are only sampling, okay if I am interested in partitioning the data I can just look at some random permutation of thicknesses right and then just chop the data at some points right so that is like sampling with sampling without replacement but I have exhausted the entire sample space.

So if I am going to do that for sure right then it can do it efficiently right then that will actually has lesser overhead than sampling with replacement but why the sampling with replacement have lesser overhead than without replacement when I do it with replaced with replacement I do

not have to keep track of what I have already sampled from doing without replacement I actually have to remove it or allowed to have some bit that sit there.

Let us say this string has been visited so how now how do I change my sampling distribution so that I avoid those data points which I have already sampled yeah, what is it now you can you can pre-roll your sampling yeah so if you are going to just sample one at a time then it is a then it is a problem right if you can pre doll your sampling that is essentially what both of you are saying that you can what I am saying is you a priori generate all the random seeds you want right just to the permutation and then just keep popping off the whatever is the ID at the topmost is a way of implementing that can be pretty efficient.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in