Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 11

Lecture – 51

Hierarchical Classification with Example in R

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture we had discussed about how are you going to do the hierarchical cluster analysis or how are you going to implement different types of hierarchical methods for clustering in the R software. And we have taken a very small data set and the objective was to explain you what is really happening behind the curtain that means behind the software. Whatever the software outcome is there how it is corresponding to the input data set. Now in this lecture we are going to extend the same commands to a better data set and for that we are going to use a built-in data set on the USArrest which is available inside the R software right. So, there are 50 observations on 4 variables still I have not taken a very big data sets because it can my graphics will become quite clumsy and it will be very difficult for you to understand but this is a quite popular data set in the R software.

So, I will try to use this data set at as many as possible places in the further lecture also. So, now in this lecture I am going to do two important things one whatever we had repeated in the last lecture the same steps I will try to repeat in the new data set. Beside those things I will try to give you some more aspects that if you try to opt different types of clustering schemes then how they are going to be compared how they look like how are you going to make a final decision. Beside those things there are several graphical ways which helps us in taking a better decision for the clustering.

So, all those things I will try to explain you I will try to introduce it and then I will try to implement it so that you can understand them. So, let us begin about this lecture and try to understand the hierarchical classification with some examples in the R software. So, I am going to take here only one example which is based on the data set which is available

inside the R software which is USArrests. This is the name of the data set USA in uppercase alphabet a-r-r-e-s-t-s in lowercase alphabet. So, actually this is the data about some violent crime rates by US states and it contains some statistics in a risk per 100,000 for assault, murder and rape in each of the 50 states in 1973 and also given in the percentage of the population living in the United States.

So, essentially, I am not interested in this story, but what I want to understand is that we have here 4 variables and, on each variable, we have 50 observations on these 4 variables which are indicated by here murder, assault, adult population and rape. So, all this data is there and here are the details which I have simply taken from the R software. Anyway, this is not a big deal for us now. We would like to just give it a name so that I can use it easily. So, I try to give this data as USArrests or say USArrests as data USA, right like this and you can see here this is the screenshot of the data set.

I will try to show you on the R console, but you can believe at the moment that you are going to get the same thing. Now, in order to do the clustering, we would like to scale the data so that the data is unit free. For that we are going to use the command scale and if you try to recall we had learnt about this that scale x, center = TRUE, scale = TRUE. We use the scale data. So, I will say scale this data USA and using the command scale and then I will give it a new name.

Data USA scale. So, you can understand it easily in the further slides. Now after that as I told you in the last lecture, we are going to use these three packages cluster, factoextra and dendextend. So, we try to install it and then upload it although we did it in the last lecture also. So, now, first try to use the command here hclust and we will try to use these agglomerative methods with different options like as complete, average, single, var etc.

So, we try to compute the dissimilarity values with this option dist and then feed these values in the hclust and then we specify these methods, one of these methods and then we can plot the data and we can see that how these methods are giving us different outcomes. So, first we try to create here the methods. So, for that I use here the command here dist, the same data set data USA scaled and method I am using here Euclidean and this name I am giving it as a distusascaled. This is how I am giving the name and then I try to use the command hclust for the hierarchical clustering and I use here the method = complete. So, this is the clustering using the complete linkage and whatsoever is the outcome I am trying to store it into hc, hierarchical clustering and then l that means hcl and then complin c o m p l i n (hclcomplin).

Right, so means h means hierarchical, cl mean cluster, comp is complete, and lin is for linkage. So, it is easy for us to refer later on. Now, we try to plot this using the command plot, but the same command what we use in the last lecture. So, we try to plot this outcome here and you can see here that this is the distance matrix of the scale data. You can see here this there are different states, Alaska, Arizona, Arkansas, etcetera on the rows and column and these are the different values of the distance using the complete linkage.

For example, this 2.717 is corresponding to the distance between Arkansas and Arizona right and so on. So, this is how you have to see it. Well, anyway if you try to then try to plot it then in the plot there I have here two commands cex and here hang. So, this cex is actually a number which is indicating the amount by which plotting text and symbol should be scaled related to the default.

One is equal to default 1.5 is 50 percent larger, 0.5 is 50 percent smaller and so on. So, you can means choose a corrective value and I would say that you should experiment with the same data set by giving different values of cex. Similarly, it is here hang. So, hang is the fraction of the plot height by which label should hang below the rest of the plot. This I will try to show you. A negative value will cause the labels to hang down from 0. So, now let us try to see that if you try to create this thing. So, you can see here first.

This is here the names of the state j have been obtained by the using command here this hang. And now if you try to see here this is the ending one which is used which is needed by using the method of complete linkage. You can see here there is a structure and you would like to later on make here a line like this one or this one and using this dendrogram you would like to classify them into different clusters. So, in order to understand this thing in a better way, we can first try to understand what is the agglomerative coefficient. So, you can recall that agglomerative coefficients are trying to give us the degree that how clusters are, how the observations are clustered.

So, for that we are going to use the package here cluster and then we are going to now here the use here the package agnes. Agnes is now the similar thing the way you have done it with the hclust. The similar thing you have to do with this new command agnes. My objective is this on the same data set I want to show you that how the result differs and how much they are same. So, for that I am using here this cluster package.

So, I try to install it and load it and then whatsoever be the my data that was the distance matrix based on the USA data which was scaled the same data is going to use here method = complete that means the complete linkage method now I am going to use with the agglomerative command agnes and whatsoever is my outcome this I am trying to save here under this name which is here hclust and then comp and then name. So, hierarchical clustering with complete linkage with agnes. So, this is how you can understand and based on that this will be the outcome here that you had seen last time that based on that we are trying to find out the relative coefficient by giving the command here dollar ac with this outcome. So, this comes out to be a 0.85 and this is here the screenshot.

So, you can see here that this agglomerative coefficient is close to 1. So, it is indicating that it is a strong clustering structure. So, this is visible from here 0.85. Now if I try to create here the but now this dendrogram is obtained here using the command agnes earlier this dendrogram this was based on the command here hclust that you can see here.

Now, you can just try to see what is the difference between the two. I am not going into that discussion that what is good what is bad or what is better, but you need to just understand it. My objective is to explain you how to work with the dendrogram. So, now this method gives us the better clustering in the complete linkage average clustering or ward clustering. So, for that the rule is very simple try to compute the agglomerative coefficient for each of the method and check the higher the value of agglomerative coefficient the stronger the clustering is.

So, let me try to for example, that what will happen whether the package and then I will try to compare them. If I now consider the adhesive hierarchical method using the command here dyna. So, you can see here using the command this dendrogram is like this like this one. So, now you have here three types of this dendrograms and then this is here the screenshot I will try to show you on the R software, but anyway. Now I try to compute here this different type of efficient using different methods.

So, I try to use here the command agnes same data assist with this was the distance matrix of a scale data. Then now I am trying to use here the method = complete and then I try to repeat it the same command is repeated here we say here method = average and then I try to repeat the same command here with method = single then I try to repeat it with the method here ward. And in all the cases I try to compute here the agglomerative coefficient by using the same command dollar ac in the outcome. So, this is here the outcome they are the same value same names and then I am trying to put here ac and the

same thing I am trying to do with the case number 2. So, this is case then it is case number 1, 3 and here 4.

So, case number 1 is using complete linkage, case number 2 is using average linkage, case number 3 is using single linkage and case number 4 is using ward linkage and the data is the same here that you can see here in this column. Now if you try to look at the agglomerative coefficient which you have obtained. In the case number 1 with complete linkage, it is 0.85 then for the average linking it is 0.73 for this single linkage it is 0.62 and for the ward method it is 0.934. So, you can see here out of this ward method has the highest agglomerative coefficient. So, I can say here that ward method gives a stronger hierarchical clustering compared to complete average and single method. This is how you can compute these things here.

So, you can see it is not difficult. Now if you try to obtain the dendrogram of this one, the method which you have which you are trying to now say that ward method is going to give a stronger hierarchical clustering. So, we try to create the dendrogram of this case. So, this is here like this and this is here the dendrogram. And you can see here this is based on ward method.

And now we would also like to see that instead of agglomerative method if you try to use the divisive method then what happens. So, we use the package here dyna and this also works just like a genet there is not much difference in the way it is going to be implemented. So, we try to use here the command here dyna and we try to do the clustering on the same data set data usascaled. And it is stored here in this name and can we try to find out its divisive coefficient. So, divisive coefficient can be found out using the name of this outcome followed by dollar and then dc, dc means divisive coefficient.

And this comes out to be here 0.85 into 0.85. It is showing that it is 0.85 quite close to 1. So, it is indicating that there is a strong clustering. And then we try to find out the dendrogram using the command here pl3 like this one. And you can see here now this is here the outcome and this is here the dendrogram. You can work in the way you want. I have absolutely no issue.

Now if you recall in the last lecture we had used this option this cutree. The cutree was used to identify the subgroups and by creating a horizontal line and then try to defining the number of clusters what we want. Right, the same thing I am going to now do here that we have here the difficulty with the dissimilarity matrix. So, from this using the

method here Euclidean we are trying to create here the distance dissimilarity matrix and dissimilarity matrix is coming how to be here data usascaled. The same data set on which we have created earlier the dissimilarity matrix and this data is stored here in this one distusascaled, but now it is with the method is Euclidean changed. And then we try to do the use the command here hclust and using the method here var.

So, this is the screenshot, but if you try to look at the outcome that yeah you know that the command that is here cutree. So, cutree means you have to define the command here as a tree which is tree is a unit using the command this is hclust and based on that then we have to give here a integer value here k which is the desired number of groups to be obtained and h here is a say numeric scalar or a vector with the heights where the tree should be cut. And then yeah means we try to do it. So, exactly in the same way the way we have done earlier you can see here if I try to divide the whole data into four clusters using the command hclust and hclust is hclvar then it is coming here like this if you try to create the table it has here four clusters one cluster has seven values, second cluster has twelve values, third cluster number three has nineteen value and cluster number four has twelve values. And if you try to again divide them into say six groups here then the outcome here like this these are the six clusters and these are the respective numbers of units inside these clusters.

You can see here this is the same command what we had used earlier right and this is a screenshot of the same command which I have just explained you. Now, if you want to understand this these dendrograms. So, it is helpful in the software that there are various option by which your life will become easy and it is very easy to understand. For example, we can draw a border that suppose I want to have four cluster and I want to know that what are those four clusters in or what are the observation which are going to these four clusters. So, this border here border is used to specify the colors of the border of the rectangles and these rectangles will be created around the cluster.

So, for example, if you try to use here the command here plot then if you want to have here the boxes you have to use here rect.hclust and inside the data in which you want to have the border with the number of clusters k=4 and border is from 2 to 5 right. So, yeah means I would say request you that you try to please understand more details about these options from the help menu and if you try to see here if you try to use the command here plot hclvat with cx=0.6 you get here this cluster or this dendrogram. But when you try to use here this command rect.hclust(hclward, k=4, border = 2 : 5) then you get here this type of structure. You can see here there this is a box number 1, 2, 3 and here 4.

So, these four boxes have been created. This is equivalent to what you have done earlier means you want you wanted to make here a line like this one. But now these boxes have been created so that you can know that what are the units which are belonging to these clusters right. So, you can see here this is here one cluster, this is in the second cluster, this is your third cluster and this is your fourth cluster right. So, you can see here then the understanding becomes quite easier here. And similarly, there is another option here what we call as has say fviz underscore a cluster function in the package of the cluster and this helps in the visualization of the result in a scatter plot right.

If you install this package and load it and then if you try to use the command here fviz_cluster all are under lower case alphabet and with the command here list with the data = datausascaled, cluster = subgroup 4. So, subgroup 4 if you try to see this is the data what we have stored here right. So, now this based on this data set it is trying to create here this type of scatter diagram which I would show you here in more detail it is a bigger the same diagram. But so, you can see here this is one cluster and this with the yellow there is here second cluster and this is for the 6 cluster right. You can see here this is the third subgroup, this is here the fourth subgroup, this is the fifth here and then there is here a sixth subgroup.

So, you can see here and then they are indicating here how this colors indicating right. So, this is how it becomes easier for us to work in the trying to work on it right. So, and before I try to go for the next function let me try to show you these things on the R command also. So, that you are more comfortable and confident. So, you can see here this is your here datausa, datausa you can see here this is the data here right.

This is the built-in data set. So, you do not have to worry for these things and then you have to you have to scale it. So, I try to scale this data and it is saved here as a data usascaled right. So, if you try to see here this data will look like here this. But now it is scaled.

Then I try to remove these libraries. So, that later on I do not have to do anything and if you remember in the last lecture, we had used these libraries. So, there is absolutely no issue and now I am trying to create these and plotting of the dendrogram here. So, if you try to use this command in the R software it will look like this right. So, this is the dendrogram which you have obtained here and if you try to find out here data usascaled matrix how it looks like you can see here this is the long data set. But anyway, I have to give you here this screenshot with a very small font like as here you can see right.

And now I try to use here the this is dendrogram and for that I try to create here this dendrogram you can see here. If you want to recreate it you can use the command here and you can see here. So, this is the dendrogram using the method complete right and this is your here this is what I wanted to show you. Now if I try the agnes package you can see here this command here and is on agnes is like this and its outcome here is like this and the agglomerative coefficient here is 0.8531583 which I can obtain here directly by writing here command ac.

So, you can see here this is my here this agglomerative coefficient which is indicating that the clustering is good right and then if I try to create here this dendrogram. So, let me close it otherwise yeah you can see here now this new dendrogram is created it is the data set based on agnes package right and then you can see here that I try to compute here all this different type of agglomerative coefficient here this is based on method equals to complete this is based on average linkage method and then this is based on say this single linkage method and now this finally this is based on this thing. So, you can see here this is based on Ward method.

So, you can see here out of these four values 0.85, 0.73, 0.62 and 0.93 the agglomerative coefficient based on Ward method is giving you the highest agglomerative coefficient right. So, this we can see here and now similarly you can compute this thing and now I come to here this the command here dyna on the same data set. So, first I try to create here a distance matrix here like this one you can see here. So, if you want to have a look it will look like almost the similar like this thing because it is a big file 50 observation.

So, that is why it is not clearly coming on my screen, but you can try and you can see and now if you want to have here this here it is a divisive constant. So, this I can find out here as a dollar DC you can see here this is 0.851. So, this is here 0.851 and if you want to create here a dendrogram and if you come like this you can see here.

So, this is the dendrogram based on the under the under the dendrogram thickness and now it is this dyna right. So, if you can see here that you are getting here now you have to see which is what is doing right, but if I try to create here this dissimilarity matrix and we try to work only with the ward method you can see here this is I am trying to just store it. So, that there is no confusion and now I will try to means create here the wards method and try to store its outcome here right and then I will try to create here this cut ray. So, if you try to see here, I am trying to create here subgroup and then I will try to made it for.

So, right. So, now if you try to make it here that command here table subgroup 4 and say table subgroup 6 you can see here it is giving you 4 and 6 cluster respectively right and yeah and if you want to have here plot command on this data set this is obtained here like this and if you want to have the if you want to create the boxes around this dendrogram in the clusters you can use here the command here this one and you can see here now this boxes have been created right. So, this is how you can work very easily and now means yeah and if you want to work here with the so you can see here that it is created here like this for subgroup 4 there are 4 subgroups and if you and if you want to have it for the 6 groups. So, you can see make it here the data with the 6 subgroups and you can see here now this there are 6 clusters. So, you can see on the screen it looks more clear and if I try to increase the size possibly it will look better you can see here. So, you can see here it is not very difficult to work with cluster analysis when you are really working in the.

So, only thing is this you have to understand and for this means you have to know the basic concepts. Now, we come to another function there is another function for E cluster right. So, I would also like to show you how it works and what really happens. So, the command is the in the use of the command is similar to what we have done in this other command. For example, if you want to use it on the same data set which is datausascaled then the command here is e cluster eclust and then you are trying to take care this whatever is your hclust that you have taken earlier k = 3 method = complete graph = FALSE and it will give you here this type of outcome right.

But anyway, let me try to go here better. So, now if I try to what is ever it be the outcome this, I am trying to store in the dataelust or I try to give the command fviz_dend on this data which we have obtained here then rectangular = true show labels = false and you can see here right it is giving you here this type of graph. So, now and in case if you want to make it here if you want to use here this option is hang = - 1. So, you can see here that in the earlier case here in this part there are no names no names of the state, but if you try to use here the hang command you can see here these names have come here because of this hang command right. So, this is how you can see here and then once again you can use the command with agnes and diana and we try to use the different clusters using this agnes command and diana command. So, for this we try to create 4 groups like agnes the same data set data usa scale method = word and cutree as into = 4 and the same thing I try to do with the diana that is cluster or say hcl_diana with the name of the hcl and I say with the diana command and then I try to use that cutri exactly in the same way and if you try to see here this will be the screenshot, but if you try to compare here you can see here that they are not actually the same they are trying to give you different values.

For example, if you try to see here California is in cluster number 2 now it is in cluster number 3. So, the use of agnes and diana you can see they are different, but this is what you have to practice and you have to understand what is really happening I can only give you here this is fundamental that what will happen. So, up to now we have this hierarchical clustering using dendrograms visually and if we want to suppose if I want to compare 2 dendrograms. So, I can plot them side by side and for that there is a function tanglegram this function plots 2 dendrograms side by side with their labels connected by lines. If you want to compare that with the outcomes using complete linkage with the Ward's method.

So, this will create a visual comparison of the 2 dendrograms obtained from the complete linkage and Ward's method. So, how to get it done? So, what we try to do here this is the same command what we have used earlier. This is the data that we have obtained here this is the dissimilarity matrix based on datausascale with method=Euclidean. Then we are trying to obtain the clustering with respect to Ward's method under the name hclvar the same command which I used earlier hclust and then I try to use the same command hclward with the complete linkage method and now I am trying to create here this hclward here like here dendward = as.dendogram(hclward). This I have to do so that I can put 2 dendrograms together and then I for this one for the second one hclcomplink I am trying to create similar dendrogram which is dendcomplink = as.dendogram() and the name of this variable.

Now I try to create the 2 dendogram side by side and for that I have to use the package dendextend. So, I already have installed it and so I try to use the library and then I use the command here dendogram inside the parenthesis try to give here both the dendrograms k-type which you want to plot side by side and if you try to do it, it will here look like this. Yeah, this as I told you in the last lecture also that when you are dealing with the large number of data then the graphics, maritime they become very clumsy but then definitely if you try to focus on it try to understand it you will understand it very clearly. You just have to spend couple of minutes only. So, you can see here this is one dendrogram and this is here another dendrogram and then they are trying to say that how something is related to somewhere like this.

So, and just to understand that whether it is that how it is working. So, let me try to take an extreme example that I try to create this dendrogram with the same data set. So, Denver and Denver. So, obviously if you try to see both the data sets are same and so outcome will look like this. You can see here there are only straight lines here that everything is matching with everyone whereas on this side if the lines are crossing that

means there is a difference in the way you have obtained the two clustering. In this case both the methods both the are giving the same outcome so the lines are simply horizontal.

But if these lines become here like this or like this that means there is a deviation. So, this output actually displays the unique nodes with the combination of labels item not present in the other tree and they are highlighted with dashed lines. So, the quality of the alignment of the two trees can be measured using the function entanglement. Definitely you have seen that here in this case there is a complete agreement between the two methods of clustering.

Whereas in this case you can see that there is some agreement some disagreement. So, how to quantify it? For that we use the function here entanglement and this is a measure between 1 and 0 and if the value is 1 that means there is a full entanglement and if the value is 0 that means there is no entanglement. So, a lower entanglement coefficient corresponds to a good alignment. So, this is how we try to do it here. For example, if you try to see here that the dendogram which we had created here this one where I have taken two different dendograms if I try to compute its entanglement this is coming out to be 0.855. So, this is how you can understand it numerically also. And now if you want to use the Elbow method. So, we just need to change the one of the argument in the function fviz_nbclust. So, if I try to just I just try to change it say if 1 = hcut and method = slot then you can see here the same command which we have used in the last lecture. So, if we can create a Elbow here and based on that we can also find how to compute that how many clusters are needed. So, I try to use here the same data and then I try to add here 1 and method = select and we get here this type of curve which is clearly visible here.

So, you can see here that we have the number of here k s number of clusters means 1 cluster, 2 cluster, 3 cluster and then this is the average select width. So, you can see here this is how you can say that we are going to find the average number of clusters. So, let me try to show you these commands in the R software also.

So, that you are more comfortable. So, I try to use this first here eclust. So, this will be here right this is my here outcome and now after this I try to create here this type of the dendogram using fviz. So, here like this you can see here this is here the dendogram which I have shown you here right. Similarly, if I try to make it here hang = - 1. So, you can see here there are no names, but there is a blank space where I am trying to move my cursor, but if I try to use here this command you can see here that these names have appeared here right. So, this is what I was trying to explain you and then I try cut these methods into two parts using here.

So, this is the eglust method which I have shown here and this here you can see here this is here cut real like this. So, you can see here this data I already have shared with you this is the screenshot and yeah if you try to increase it 6 this will come again like this one. So, now I try to create here this tanglogram. So, if I try to just copy these commands now you have understood what I am what I have done. So, without wasting much time I will simply try to copy it and then you can see here this tanglogram has appeared.

So, now it is your capability that how you want to understand it and to make you understand that if you try to take here the same data set for the dendrogram here you have taken then word and then comp link, but if I try to take here both then word and then word you can see here that this is a curve which is a straight line right. And then after that similarly means you can compute here this entanglement here right if you try to see here. So, and then I try to compute here in entanglement. So, this is coming out over here like this right. So, you can see here that it is not very difficult and then if you try to see here if you want to have here this Elbow method if you want to use you can create here this curve here like this right.

So, you can see here it is not very difficult to create this graphics and to understand this outcome. So, with this we come to an end to this lecture and you can see that well this lecture has most of the component which are related to the outcome and I have told you couple of things more which I had not covered in my theory part. But as I told you this clustering is basically a technique which is based on computation. And theoretically I can define many things, but when we are trying to do it numerically, they have to be understood that how the software is going to explain them or how the software is going to represent them.

There are different ways by which the theoretical results can be expressed in a graphical way. For example, I have shown you here, but I am not saying that I have covered all the topics in the cluster analysis. There are many more things using different types of packages you can create more type of more different types of graphs which can be more informative. And if you want to dig out some particular type of information then you can choose the appropriate graph and then it will give you a much clearer information because you have to understand you are working in the sea of data. It is so huge that it is practically impossible for you to understand the data by observing it by looking at it. So, you have to finally depend on these statistical methods and based on this software outcome you have to take a final call.

So, definitely theoretical background is very important to understand what software is trying to do, what is happening behind the curtain, but your practice is also very important. The more you work with this type of dataset, better you will understand. I would suggest you that instead of taking a built-in dataset, why do not you create your own dataset and then try to see whether this clustering can be done or not. I will say take only two types of variables or say values 0 and 1 and try to see or say two types of value 1 and 100 and try to see what is happening in the group. Try to take 500 values and try to just create the clusters and then try to see how the software is working.

Out of these different ways to handle the problem which method is more appropriate for a given dataset. This is what you have to learn which I cannot make you learn but only your practice will make you learn. So, I request you that you try to practice it and I will stop here with the cluster analysis and but it does not mean that the topics have ended. There are many more topics and I would request you please try to have a look in the books, in the software and try to explore more.

So, you try to practice it and I will see you in the next lecture with a new topic. Till then, goodbye. Thank you.