Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 11

Lecture – 49

Hierarchical Classification

Hello friend, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture we initiated our discussion on Cluster Analysis. And we had talked about different small topics to gain some idea that what are the different concept behind using the cluster analysis. So, cluster analysis is essentially used when you have a population and you want to divide them into different cluster, different groups so that they are within homogeneous. Now, the question comes here how are you going to do it? We have discussed different types of concepts now those concepts have to be implemented in such a way so that you can create this clusters. So, if I try to take say two examples and try to understand it how it can be done.

In the first example I consider suppose there are 5 units which have to be classified into different clusters. So, in this example we are trying to create 5 clusters 5 units 5 clusters. So, every cluster has only 1 unit. Now I try to see that which of the units are similar suppose unit number 1 and 3 are similar.

So, we try to club them together. So, now in the second step we have 1 cluster containing the unit number 1 and 3 and say another is 2, 4 and 5 they are in 1 unit in 1 cluster each. Now in the next step I see that 4 and 5 they are also closer to each other. So, I try to make 4 and 5 together and now I try to see that we have here 3 cluster 1, 2 in 1 cluster 3 in 1 cluster and 4, 5 in 1 cluster. Now in the next step I try to see that 1, 2 and 4, 5 they are also closer.

So, now 1, 2, 4, 5 becomes 1 cluster and 3 is separate and in the final stage at the most I can have only 1 cluster of containing all the units 1, 2, 3, 4, 5. So, in this case what we have done we have first taken the cluster of size 1 and then we have started adding them step by step. And opposite of this will be that we try to make first the 1 cluster of all the units suppose if I take 5 units. So, in the first step I try to create the cluster of size 5. So, I will have only 1 cluster.

Now I try to divide them into 2 cluster, 3 cluster, 4 cluster and at the most then I will have only 5 cluster. So, in case if you try to see in first approach I am trying to add and in the second approach I am trying to divide. And this process is being done in steps. Every step will have different types of clusters, different number of cluster and the units will be allocated to different clusters. So, these are called hierarchical methods.

So, using these hierarchical methods of clustering we can have these types of approaches in which we can construct the clusters. So, now what are the concept behind this hierarchical classification and how to get it done and what are the different concepts which are needed to do it this is what we are going to discuss in this lecture. The next question that how to implement them in the R software that we are going to take up in the next lecture. So, let us try to begin this lecture and try to understand the basic concept of the hierarchical classification, hierarchical clustering. Okay, so now we are continuing with our topic on cluster analysis and we are going to talk about hierarchical classification in this lecture.

So, in this type of classification the data are not partitioned in the single step. Instead, classification consists of a series of partition which may run from a single cluster containing all the clusters to the number of clusters each containing a single individual. Right, if you have say this unit number 1, 2, 3, 4, 5. So, in the next step you try to suppose make 1, 2 and 3 in one cluster, 4 and 5 in different cluster and finally if you follow these steps finally you will have here one cluster of all the units 1, 2, 3, 4, 5. The second option is this you have one cluster of 1, 2, 3, 4, 5 then you try to divide them into suppose 1 and 2 then 3, 4 and 5.

Then again in the next step you can try to divide you can try to add more units in one cluster and finally you will have say 1, 2, 3, 4 and here 5. Right, so this is what I meant when I wrote this sentence. Right, so in this type of classification you have seen that there are basically two approaches. One is agglomerative approach and say another is divisive method. Means agglomerative as the name suggests that you are trying to accumulate and divisive method means you are trying to divide.

Right, so this hierarchical classification can be subdivided into two approaches. In these agglomerative methods they proceed by a series of successive fusion of the n individual into groups and they are shortly called as AGNES that is agglomerative investing. Right, and in the divisive method the divisive methods separate the n individual successfully into final grouping. Right and this is called as DIANA which is the short form of divisive analysis. So, these are the mainly two approaches to consider this type of clustering.

So, if you try to understand then suppose if I try to first consider the agglomerative methods that how do they work. So, let me denote here by here capital Pn this is the stage where the n individual which consist of n single member cluster which is also called as roots. And suppose P1 is indicating that it consists of a single group containing all the n members and this is called actually as a leaf. So, now when we are trying to consider the agglomerative methods then they produce a series of partition of the data from Pn to Pn - 1, Pn - 1 to Pn - 2 and finally up to P1. Right, so how it happens? So, in the first step we start creating say n clusters.

So, we have n units so there are n cluster. So, that means every cluster will have only one unit. Right, so each of the size containing single element and we indicate them by here C1, C2, Cn. Right, C1, C2, Cn each has one unit. Then we have we try to find out the nearest pair of the different clusters and we call them suppose here Ci and Cj.

Suppose Ci and Cj they have got the closest to each other or nearest to each other now the question is how you can get it done. So, if you try to recall in the last lecture we had talked about different ways by which you can measure the distances. Right, for example Euclidean distance, Canberra distance, metric distance etc. So, now you can use any of those distance to find out this nearest pair of this cluster. Now I try to now make one cluster of Ci and Cj.

So, we delete Cj and decrease the number of clusters by 1. Right, because out of C1, C2, Cn this Cj has now combined with Ci. So, the number of clusters will be reduced by 1. So, and this process is continued and at each step we try to or we fuse individual or groups which are closest and actually they are called here as a node. Right, now the question is that there are different methods, different ways of finding out the closeness, different ways to find out the nearest pair of clusters.

So, different methods based on different similarity measures and distances can be evolved and they may or may not give you the different outcome. This is very important for you to keep in mind and you will see that when we try to implement it in the R software then the package ask you for this option that which of the measure you want to use. So, this is important for you to understand and keep in mind. Right, and there are some limitations. For example, in the division methods or in the agglomerative methods the divisions or the fusions are irrecoverable.

That means once 2 cluster has been joined then in the next step they cannot be separated or once separated, they cannot be joined. So, this approach is only one way. Either you are trying to fuse different clusters together or you are trying to divide different clusters they are not reversible. So, a decision is needed for the investigator to decide to stop at a particular stop with the optimal number of clusters. Right, suppose if you start with one cluster having 1, 2, 3, 4, 5 and now finally you will have here 5 clusters say 1, 2, 3, 4, 5 for example or the reverse way.

So, there will be different steps here. Now somewhere the investigator has to stop because having only one unit in one cluster is not advisable then why should you do the clustering. You want to make the groups homogeneous or you want to do the cluster homogeneous. Right, so that is why there is a concept or there are ways to find out such optimal number of clusters that we are going to discuss and in this case the concept of dendrogram helps us. So, this classification means hierarchical classification either agglomerative methods or divisible method-based classifications they can be represented by a dendrogram.

What is this dendrogram? Dendrogram actually illustrates the division or fusion at each successive stage of classification. You can recall that in the last lecture towards the end I had made this type of graphic. So, that was actually dendrogram and if you try to understand how it looks like then it will be easier for you to interpret the result from the software. Well, these dendrogram they will be created by the software. Yes, you can do it manually but it is very time consuming and tedious job if your data is large.

So, you can just try to understand first the agglomerative method. But first we try to understand what are the different notations here. Suppose I try to take here 5 units which are indicated by here A, B, C, D and E and they have been contained in one cluster. These clusters are indicated by these rectangles. So, these clusters they are called here as a leaf.

And finally, if you try to see towards the end here there is only one cluster which is containing here all the units A, B, C, D and E this is called here as a root. So, these are the different standard terminologies in the clustered analysis. Now we try to move ahead and we first try to understand the agglomerative method. So, if you try to see here 0, 1, 2, 3, 4 they are trying to indicate the stages something like P0, P1, P2, P3 that we have indicated P4. And the same stages have also been given in the bottom but I will try to explain to you later.

First let me try to understand here the agglomerative method. So, now if you try to look at here this type. Now I am using a different colour of pen. If you try to see here this A and B suppose using any distance major they have been found that they are close to each other. So, we try to merge them into one cluster.

And at the same stage we also find that clusters D and E the units in the cluster containing the unit D and E they are also connected to each other. So, we try to merge them together into one cluster that is D, E and C, E is not similar to any one of them. So, now if you try to see at this stage, we have here cluster number 1, cluster number here 2 and C will remain here as a cluster number 3. Now in the next slide what I try to do here that I try to come to this stage let me use here a different colour pen this green colour. So, now you can see here that in the next stage this C and this second cluster this one they are closest to each other and we try to merge them together.

And this A B will continue as such there is absolutely no issue. Then in the next stage they remain separated there is no issue. So, you can see here at this stage we have cluster number 1 and cluster number 2 and here at the next point there is no there is no change. Now in the last stage if you try to see if means everything is fine and there is no option left then all this cluster A B and this cluster with here C D E they are merged together and we have only here one cluster. So, if you try to see when you are going from left to right like this then gradually the units in the clusters are actually fused together and the numbers of clusters are decreasing from say here, we have here total 5 clusters and here we have only here one cluster.

So, if you try to see you are trying to fuse different clusters together and so this is called as agglomerative method. Now if you try to think about the divisive method then they are just opposite to it. So, let us try to understand it from this clean graphic once again. Now if you try to look at this bottom so we have here we are trying to indicate the divisive methods right. So, it is just opposite to the agglomerative method if you go in this direction.

So, if you see here at this P0 stage you have only here one cluster one cluster which is containing all the units right. Now we move forward and we try to measure the distance so in the second stage we do not find anything but after that in the next stage what we try to see here that C, D and E they are close to each other. So, we try to bring this C, D, E into here this group and the remaining A B they will continue. Now in the next stage then what we try to do here that now C, D, E we try to look at this cluster and suppose this is divided into two clusters one containing here D and E and say another containing here C right. You can see here C will continue here and then A B is now continue here as A B.

Now in the next step you can see here this A B is divided into A and B C will continue as C and D and E they are divided into D and E. So, if you try to see here now towards the end you have here five clusters and in the beginning you have only here one cluster. So, you are trying to divide the one cluster into different segments different groups and the maximum number of clusters can be five because there are five units in the one cluster here A, B, C, D, E. So, now you have to see depending on your need and requirement where you want to make a break line. For example, whether you want to have only C, D, E and A B together or you want to have A B in one C into in another and D E in say another cluster right.

So, at this P3 stage you will have here three clusters at right and so on. So, this is a break line. So, now you have to so this is a decision which we have to make. Well, I have taken here a very simple example to explain you. Rest you will see that these things are automatically created in the software and whether you want to have three cluster, four cluster or five cluster this is your choice this is your decision and based on that these types of lines will be created in the dendrogram.

The only difference between the default method of creating the dendrogram in the software and here is that here I have taken the horizontal but in software you will see this will become a vertical like this one and so on. But anyway, you will also have an option in the software to make it horizontal. So, anyway so do not get confused when I try to show you this dendrogram in the R software right because I have created only the vertical ones which are easy to understand actually right. Now the next question comes here that how are you going to say decide that how many clusters are needed right. How many groups should be there? So, what we can do here as I shown you here that you need to cut the dendrogram at the place where the partition produces the best fit of the data right.

You have to see given the conditions of the data and the requirements means everything together that where you can make a line to cut the dendrogram and then you can decide that which of the units are going to which of the clusters. So, informally examine the difference between fusion levels in the dendrogram. That we can see that okay that the fusions level has to be in such a way such that within clusters the unit should be more or less more similar right and in order to do it yeah manually it is difficult but different optimization methods are available in the literature and they have been implemented in the software also. Although I can share with you that I am trying to give you here the basic idea and then there are different methods which can be implemented and then you can you will see in the next lecture that just by using the different options in the R software you can easily do it. So, now when you are trying to work with this dendrogram so you have to keep in mind that each leaf in the dendrogram corresponds to one observation which is displayed there right.

You can see here that these are your hair leaves right. As we move up the tree observation that are similar to each other are combined into branches which are themselves used at higher height. You can see here in this graph these are here actually branches these lines which I am trying to make here circle these are the branches. So, they are helped to and they are helping in combining the observation or helping in fusing the observation in clusters. The height of the fusion provided on the vertical axis indicates the similarity or dissimilarity between the two observations right.

For example, if you try to see here in this one you can see here that at this level the clusters AB and DE they are quite similar to each other right. Wherein in the first stage all ABCDE they are similar to each other right. And so, the height of this fusion is going to help us in taking a appropriate decision. The higher the height of the fusion the less similar the observations are.

This is how you have to keep in mind. Well, I am trying to explain you here all these basic concepts and in the next lecture I will try to take say couple of examples where I will try to take the big data set also. There you will see that the dendrogram structure is not is quite clumsy there are many clusters and then so on. Then the difference in the heights may not be clearly visible on your computer screen. But you have to understand the concept that how are you going to take the decision. And these decisions will be varying from one person to another person.

So, hopefully I will be showing you the dendrogram but how to take the decision whether 5 cluster or 7 cluster that is the decision that will be dependent on the experimenter or you people. I will try to show you that how to create 5 and how to create 7 clusters right. So, with this objective you please try to understand this lecture right. So, the conclusion about the proximity of two observation can be drawn only based on the height where branches containing those two observations first are fused right. So, that is what we have done in these two this dendrograms for illuminative and divisive methods.

And the proximity of the two observations along the horizontal axis cannot be used as a criteria for their similarity. Please keep this in mind. And the height of the cut to the dendrogram controls the number of clusters obtained and it plays the same role as in the k in the k means clustering. k means clustering we are going to talk about it. So, we will see that too that this number is going to play a similar role what the value of k is going to be right.

For example if you try to see this will be a very clean dendrogram that is produced in the software right. You can see here now if you try to see here these are here the leaves you can see here right and go on means finally, if you try to see here there is only here one cluster containing all the units. But now once you go down then it is divided into this cluster at the last stage and then so on it is divided into fourth stage here like this and so on. So, this is what you have to understand and you can see here the difference these types of heights. This is what I was talking about this difference in the heights it will give you the idea about the similarity or dissimilarity.

But definitely as I said this is a very neat and clean dendrogram. So, but in but when you have large number of observations that will be little bit clumsy right. So, if you try to see here if you want to decide in this dendrogram that how many clusters you want to have four clusters or say six clusters. So, what you can do here if you want to have a only four cluster then you can cut the line here and I am going to put this green dotted and then you will see here that you will have here four cluster one corresponding to this two three and here and therefore. On the other hand, if you want to have six cluster then you have to cut the dendrogram at this place and then you will have here six cluster.

But definitely I would say that different similarity measures and distance measure produce different clusters. So, there is no unique choice if you try to take one measure and if you want to have four clusters and if you take the second measure and if you want to have the four clusters then these clusters may be different. And even if you try to make this line if you fix this line somewhere and if you try to create the dendrogram with two different measures then the same line on the two different measures will give you different clusters in the dendrogram. So, this is what you have to keep in mind. So, for example if you try to see this is only an example here that I have taken some data that we are going to use in the R software.

So, I am trying to use here different types of similarity and distance measure to produce the clusters and you can see here there is no unique choice. For example, this cluster is based on the complete linkage if you try to recall we had done it. This is made means that second one is based on a single linkage that we discussed in the last lecture. Similarly, this another two is based on average linkage and Watts method.

So, you can see here clearly that these are quite different. So, different similarity measures and different distance measures will produce different clusters. This is what I wanted to inform you. Now when you are trying to understand the clustering you are trying to measure the distance between different units. So, now you have to understand under what type of condition you can say that whether the two groups are closely related or two groups are well separated. And whenever you are trying to merge the groups or say divide the groups then the tendency to cluster together will also be changing.

And when you are trying to cluster which are quite similar to each other that means that tendency to get closer together is relatively low. So, that means all the units are becoming more similar within a cluster then this is called as screening. And in this case individuals are linked by a series of intermediates. And they fail to resolve relatively distance cluster where there are a small number of individual lines between them.

And such type of points are called as hill noise. For example, if you try to look at this picture here you can see very clearly here that these points, they are very clearly trying to identify possibly a cluster. Now what about these three points? Particularly this point will go to here or this will go to here we don't know. This point will go to here or not that we don't know. So, there is a confusion. So, these two clusters they are well separated clusters say cluster number 1 and cluster number 2.

And all these points point number 3, 4, and 5 they can be called as noise points. So, obviously you would like to have a good clustering where the noise points are the minimum. Now in order to know that that how many clusters should be there because we are talking of the variability that we want to have those clusters which are more similar in size. So, we can create different types of clusters and then we can compute this their variability and based on that we can find the optimal number of clusters where we feel that the variability is relatively lower.

Yeah, it cannot be made to 0. So, in order to know the optimum number of clusters we have a method what is called as elbow method. What is the elbow? Elbow is in your hand. So, that is the same elbow if you try to see this is your ear hand and if you try to see here this is your ear elbow. So, this is your ear hand. So, if you try to see here this shape is usually like this one or if you try to revert it will be something like this but elbow is the point where you are trying to make a turn.

So, there are several approaches to determine the optimal cluster with the hierarchical clustering. So, among them we are going to understand the concept behind the elbow method. So, in this method in the first step we try to or we run k means clustering on the data set for a range of value of k from 1 to 10. And then we will try to calculate or we will calculate the total within cluster sum of squares which is WCSS within cluster sum squares for each value of k. For example, you take k equal to 1 then try to compute WCSS1, k equal to 2 and try to compute WCSS2 and so on.

And then we try to plot a line chart or the curves of the WCSS versus k. If the line chart looks like an arm, then the elbow on the arm is the best value of the k. This is my elbow so this is value of here k. So, the elbow method does not work well especially if the data is not very much clustered. Well, every method has a limitation. But anyway, we will try to show it and you see it will look like this that I am trying to take here the number of clusters here k, 1 cluster, 2 cluster, 3 cluster, 4 cluster.

Definitely if you try to increase the number of clusters then the variability within the cluster is going to be decreasing. So, this is what is happening that if you have this from this point to this point and you can see here. Now the main challenge here is that you have to find that where is the elbow. The elbow is here or here that is where there is some confusion but if you think practically, it hardly makes much difference in most of the cases whether you have 4 clusters or 5 clusters.

Definitely elbow is not here no elbow is not here no. So, now whether you are trying to take here 3 clusters or 4 clusters or say 4 clusters or 5 clusters right this will not make much difference between the two right. And what we want? We want a small WCSS because that is the variability and this WCSS tends to decrease to 0 as we increase the number of clusters k right. And in extreme case what will happen that WCSS is equal to 0 when k is the number of data points in the data set because each cluster is of size 1 with 1 data point right. And so consequently there will be no error between it and the center of its cluster means you are simply trying to say ok say and values and data points absolutely there is no variation. If every group has only one observation whose variance will always or whose variability will always be 0.

So, that is why if you try to increase the cluster size this WCSS will increase. But definitely when you try to take only here the one cluster of all the unit then the WCSS is going to be highest but we want to strike a balance between two. So, our goal is to choose a small value of k that still has a low WCSS right. So, the location of a bend or the knee in the plot is considered as an indicator of the appropriate number of clusters. So, the elbow represents the point from where we start to have diminishing returns by increasing the value of k right.

So, that is how we try to do it and you will see that ok. Yeah, this is produced in the software so you do not have to do it manually. Then there is another method which is average skillet method. So, in this average skillet approach we try to measure the quality of a clustering and it determines how well each object lies within its cluster. So, i value of this skillet width this indicates a good clustering and in this average skillet method we compute the average skillet of the observation for different values of k. And then we try to choose the value of k such that the optimal number of clusters k is the one that maximizes the average skillet over a range of possible values of k.

And in the R software we have a function here skillet silhouette by which you can compute this width this average skillet width. So, it is not very difficult actually in the R when you try to implement it. And the skillet value for each point is a measure of how similar the point is to point in its own cluster with respect to when compared to points in other cluster. And this skillet value Si for the i-th point is defined as here say bi - ai divided by maximum value between ai and bi and it always lie between - 1 and + 1. Where this ai is the average distance from the i-th point to the other points in the same cluster as i and this bi is the minimum average distance from the i-th point to points in different clusters minimize over clusters.

And then anyway so these things are computed in the software so you do not have to worry but you have to simply understand what these values are going to indicate. So, a higher Si value indicates that i is well matched to its own cluster and is poorly matched to another cluster. And in case if most points have a high skillet value, then the clustering solution is appropriate that is what we are trying to say. And if a in case many points have a low or negative skillet value then the clustering solution might have too many or too few clusters. And one of the one criteria which we use in practice based on the select value to find any distance matrix as the clustering the clustering evaluation is as like this.

So, these are here the range of Si's. So, we try to and try to say that if the range of Si's between 0.71 to 1 this is indicating a strong structure has been found. If it is between 0.51

to 0.7 then a reasonable structure has been found. If it is between 0.26 to 0.50 the structure is weak and could be artificial. And if it is less than 0.25 that means no substantial structure has been found. So, based on the value of SI this is how you can take a correct decision or a correct interpretation.

So, you can see here based on that for example in this type of graphic you can easily find out the value of k. For example, you can see here these are the average skillet widths on the y axis and here is the number of clusters here k. So, you can see here wherever there is a change here this possibly can be the optimum number of clusters to be found. So, now if you try to understand finally what are the different steps which are required for this hierarchical clustering when we try to use any software R say.

So, we need to have a software that is decided we are going to use here R. We need to prepare the data for this hierarchical cluster analysis. We need to compute these different types of quantities for the clustering. Then we need to create the dendrograms and understand them and then we need to determine the optimal number of clusters to group the data and based on that we have to move further. So, now we come to an end to this lecture and you can see here this was a pretty important lecture for us in which we have understood what is about the hierarchical clustering because this is the clustering technique which is going to be used. Well, I have not taken here any example where I can show but my objective was that first let me prepare the background or I should give you here the definitions.

So, that in the software when I am trying to do it there should not be any problem for you to understand it. So, now my request to you is that it will be very helpful for you to understand the next lecture provided you are very clear about different types of concepts and definitions whatever I have explained you in this lecture. Because when I try to show you the software outcome then possibly, I may not have time to explain these things once again. So, I will simply use okay this is my this method and this is the value of SI, this is my elbow method etc. and then I will try to show you the outcome.

My basic objective in the next lecture will be to show you that how are you going to implement it and how are you going to interpret the outcome of the R software. So, with this request I request you once again to revise this lecture very carefully and I will see you in the next lecture till then goodbye. Thank you.