**Data Analisys and Decision Making - II**
**Prof. Raghu Nandan Sengupta**
**Department of Industrial & Management Engineering**
**Indian Institute of Technology, Kanpur**

**Lecture - 56**
**AIS**

Welcome back my dear friends and students a very good morning, good afternoon, good evening to all of you wherever you are in this part of this globe. And as you know this is the D A D M, which is Data Analisys and Decision Making-II course under N P T E L MOOC series. And this total course duration is basically for 12 weeks which is basically 30 hours, 30 hours get converted into 60 lectures because each lecture is being for half an hour. And we have already completed 11th; 11 weeks; we are going to start the last week which is the 12th week which is the 56th lecture as you can see. And after each week you do 1 assignment.

So, you have already completed 11 assignments, you will be doing the 12th one and then the final examination would be there. So, and my good name is Raghu Nandan Sengupta from the IME Department IIT, Kanpur. So, we remembered we were discussing in the 55th lecture starting that and 54th also I mentioned something, about different type of heuristic methods.

And one was basically the artificial immune system; that means how the body generates immunity system, the fighting mechanism and how they fight the pathogens or the bacterias and the viruses which come. And based on that we basically develop the artificial immune system as a heuristic methods and under that you had the negative selection, the positive selection and the clonal selection method.

Negative and positive word in the sense that in the negative one you go and try to build up a set consisting of and obviously, I did mention that one of this distance measured based on which the matching would be done to find out the best solution was in the hamming distance, there are other distance measures also which is 1 norm, l 2 norm, l infinity norm, Mahalanobis distance, maha (Refer Time: 02:24) distance and all these things are there. And in the negative and the positive selection as I said they are based on opposite principles, in the first case you develop and find out the set initially the set is empty; you fill it up with the best possible so called set of solution and in the other

method you build up the set which is also initially empty, the worst set of possible solutions.

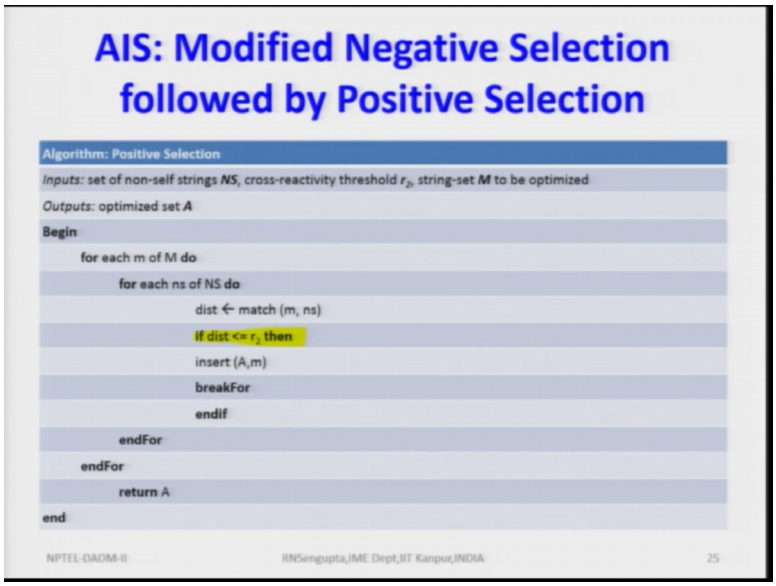
So, in the example which I was considering; I was basically trying to build up that how I can find out the bankrupt company or the sick companies, which was sick. And we have a propensity to show that they will be turning sick in the coming days or years in the future. And we took basically technically as I mentioned that historically 1960s there were lot of studies starting by Altman.

So, you have the Altman score, Zeta score, Olsen score, (Refer Time: 03:26) score. So, all these things were there, they were all based on this concept of using the statistical methods to find out, how good or bad the sickness prediction could be done. The concept of principle component analysis all this things are generally used. In when I am talking the concept of multiweighted statistical analysis based on the fact which the Altman score was developed.

Now we are trying to take the same variables; variables means the 14 such financial ratios, that you can assets the asset to total sales, then the values of sales to the values of the liabilities all these things can be taken. And one can take the stock market price also which gives a good prediction of the sickness of the company.

And then we said that we will try to basically discuss the algorithm; algorithm in positive and negative and then basically show you results about that. So, I had given you that background and based on that we will continue the 56th lecture.

(Refer Slide Time: 04:46)




So, in the positive selection method you basically set up a non-self string; non-self means self means which are maximum matching. Non-self means we are not maximum matching, they are farthest away from the set based on which we are trying to build up the set of strings based on which we will do the prediction concepts. So, inputs are the self or non-self strings, which we name as N S. We will have a cross-reactivity threshold. So, if the threshold crosses does not cross based on this we will basically match it. And we have string set of M which is to be optimized M number of values to be optimized.

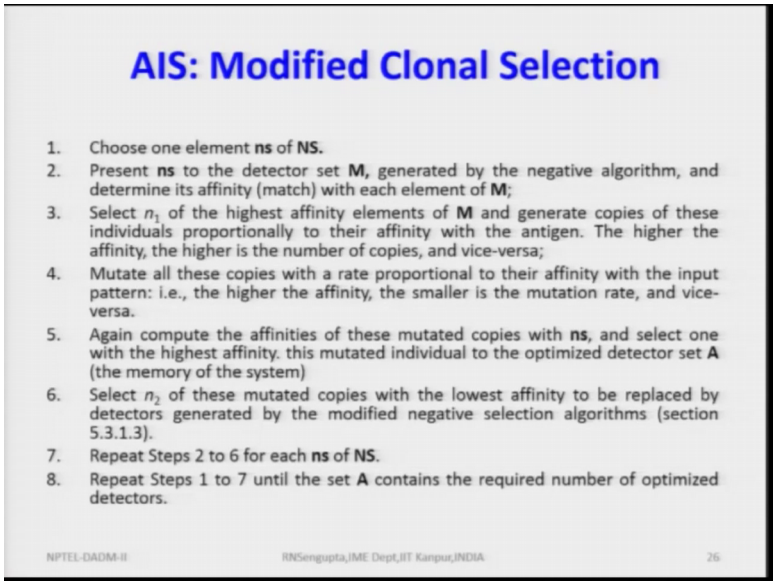
So, they would be basically normalized on a 01 scale. So, there are 14 say for example, there are 14 ratios, all these 14 ratios one by one will be matched. Based on that we can say, that how good or the bad is bad the match is. Now remember one thing you can have this threshold, which I am talking about here it is mentioned as r 2, this threshold can be each and every ratios can be matched and that difference if it is less than or more than a (Refer Time: 05:58) epsilon will activate that the yes the matching is happening or no the matching is not happening; that is 1.

Number 2 is that it will also mean that out of the 14 ratios if say for example 5 of them matching, and not matching is happening and they have crossed the threshold value in whichever sense you are saying they are less than or greater than; we will also say the matching has happened, because the ratios by themselves have to be compared and collectively we will find out how many such ratios crossed the threshold based on which

we can say or the company is heading for a bankruptcy. So, our optimized set would be A; so for each of m small m of a of capital M. That is the capital M is the total number of such string sets are there. We will do the matching and it is mentions here, if you read this, then if you see this line.

So, if the distance is less than equal to r 2; so that is the values based on which you are going. You will take that the set and put it in the positive set. So, in the positive selection, you are taking into positive set and continue till you exhaust the search.

(Refer Slide Time: 07:11)




Now, the modified clonal selection model is that you will choose one element small n s out of this no-self capital N S. So, there are many non-selfs; you will choose one. You will present n s to be the detector set in that M. So, if there is a non-self; it will trigger and the detection and non detection will occur. So, we are saying that detector set is that it will generate it by the negative algorithm, based on the negative algorithm means they are farther away.

So there is no matching. Positive means they are matching. So, we are just taking complementary part in one case matching, another case not matching which are generated by the negative algorithm and it define its affinity or the match; then affinity of the level of value of n r 1, r 2, r 3 whatever we have. Select small n 1 of the highest affinity cells from M or elements from M and generate copies of this individual proportional depending on the affinity value. So, the affinity is very high. So, you will

generate more of them, affinity is low you will generate less of them. So, it is exactly like this when the body is been attacked by germs and viruses, and consider it is a flue and flue has a certain set of say for example; pathogens and the anti bodies want to fight them.

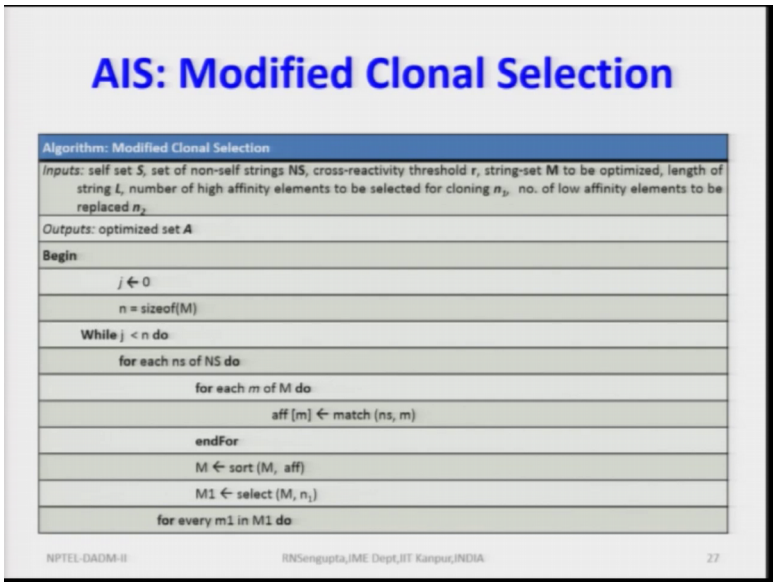
So, the fighting mechanism based on which the dovetailing happens, where they lock the lock and key concept if you remember I discussed that in the diagram concept. So, they would be generated more, because they need to find fight the flue; virus or whatever the other bacterias it can be Hepatitis B it can be pneumonia, cholera whatever it is. Consider the body initial resistance are there. I am not talking of the sense that the drug or the medicine is being given, and I am not going to that stage.

And once the higher the affinity; the higher is the number of copies generated as, I said more the attack is there for a particular type more such white blood corpuscles the fighting cells would be generated such that, they can fight those specific pathogens which are attacking. So, in the 4th step it will mutate all these copies with a rate proportional to the affinity as I said more; more you generate.

Again compute the affinities of these mutate copies with n s and select one with the highest affinity. This mutate individual to you would be basically set as the first level of memory cells. Then select n 2 and then it will be replaced by it and again you will generate the second level n 2 like as you have done n 1 and you basically repeat the steps for each and every this affinities and then you are ready to fight the pathogens which are attacking.

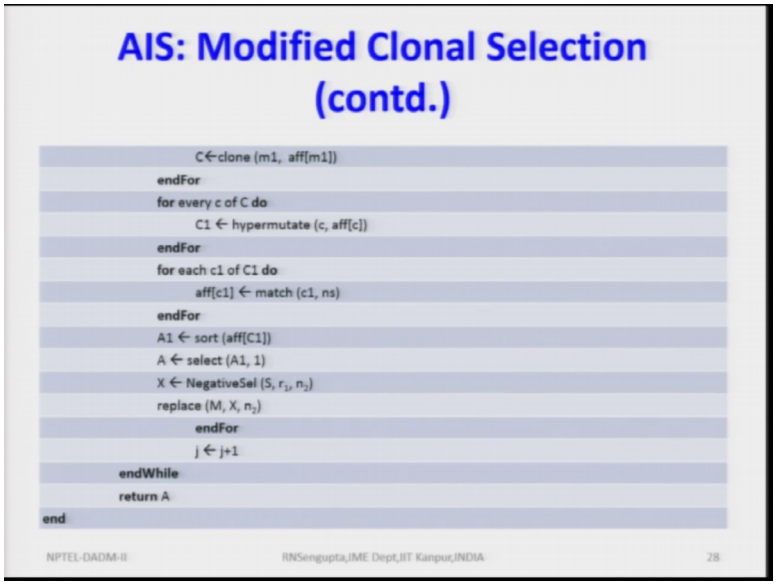
This is the clone; you clone yourself, mutate yourself; yourself means the cells mutate themselves in order to fight.

(Refer Slide Time: 10:15)




So, this same thing you take a string small n s of n n s; a threshold value is calculated. So, this threshold values would basically change as I am changing n 1, n 2, n 3. So, I take the first threshold value consider is r i. So, r i, r 1, r 2, r 3, r 4 are the threshold values. For r i; I find out n 1 clone them and set them aside. Then for r 2 affinity I again clone them find out n 2 and so on and so forth till the matching is done such that my cells are there to fight the germs. So, this is how the clonal selection occurs.

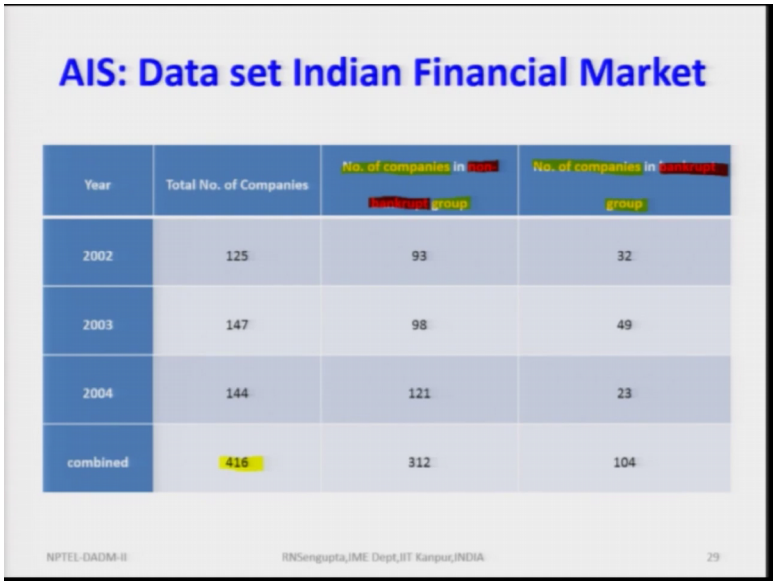
(Refer Slide Time: 10:57)



So, I am not going to the details, but the general the loop by loop, they will be basically performed such that you replicate all the capital N S number of them. Now in the clonal selection would have a complementary part, when while basically discussed that from the point of view a positive algorithm a i. So, in the sense the negative one was I will basically generate based on the negative one and in the positive one; I will basically generate on the positive one such that complementary concept of negative and positive can be combined in their respective clonal selection algorithm.

Clonal selection algorithm is just basically a subloop such that you are able to create more of these negative one or more of these positive ones depending on what selection algorithm you are trying to use, if it is negative you generate more and more negative. If it is positive you generate more and more positive. Why; I am saying that I will come when I go into the details of the matching problem.

(Refer Slide Time: 12:05)




So, here I start discussing the problem. I am not going to solve it; I will only give you the results. So, the data set is taken from the Indian Financial Market. So, I am taking a companies in 2002, 2003 and 2004. It can be done for the latest data also the, but this work was done based on the data sets such that you have two 2002, 3, 4.

And the total number of companies in the respective years 2002, 2003 and 2004 are respectively 125, 147 and 144. So, the total combined value comes out to be 4 1 6. Now out of this the what is now your question would be in 2002 what is with this 125 or in

2003 what is this 147 and similar in 2004 what is this 144. Now look carefully the second column and the third column has their heading, I will mark it with a different color; it is the number of companies in the group. What is that group? I will mark; is non- bankrupt which is important to for you to understand, it is a non-bankrupt.

So, they are not sick, they are good companies; good companies in sense based on the ratios or whatever financial factors you are going to consider. In the second set number of companies in this group, what is this group? They are in the bankrupt companies. So, I basically taken 125 set. So, this by the way; the set of companies which is bankrupt and non-bankrupt it is a sample set which I am taking. And they can be from different sectors or the same sectors. Now the number of companies which are non-bankrupt is depending on.

So, they can be two type of things. I think the companies were bankrupt were not doing well, but now they have come out from the stage where they were not performing well, and now they are healthy doing good or I can take arbitrarily some set of companies from the same sector with the similar type of ratios. So, say for example, I am considering x y z company which is bankrupt. So, I will also consider a similar type of company in the similar sector which has almost the same type of products being produced same size so on and so forth. But it is a good company.

So, I will basically try to compare apples and with apples; not apples or oranges. In the sense; if a company say for example, steel sector of a size of 100 crores is not doing well, I will consider the same type of steel sector company of in the range of 100 crores 150 or little bit less than that, less than 100 crores which is doing well.

So, similar type of number of products they are making, similar type of raw material they are trying to utilize, almost similar type of number of factories they have. Is the first set of company which is not doing well; it has three factories I will also consider. The second company which is doing good has 3 such factories number of employees and all these things are matching should match. So, based on that I have a non- bank company and a bankrupt company.

So, if you add up this 93 plus 32 comes out to be 125. Similarly 98 49 comes to 203. And this set of 204 on in 204 121 and 23. Now I remember one thing you may be thinking that what does this number is number of bankrupt companies is increasing. It is 32, 49 23

increasing and then decrease. So, you may be thinking that, they are common in the years. So, the reason is the bankrupt companies are not common. In the sense that if I consider in 2002, these 32 number of companies were bank declared bankrupt. So, obviously, any revival strategy which ever was been taken for these 32 companies does not come into our consideration for the next 2 years of our study. This 49 companies in 2003 which were bankrupt, we will consider those companies are fresh.

So, they may have been in a stage in 2002, where they were showing signs, but we did not take them. We are going to take them only as bankrupt companies the moment they fail. Any indication we are not going to consider, similar 23 number companies are bankrupt in 2004. Now in the bankrupt; non-bankrupt company you will be thinking that this number is increasing 93, 98, 121.

It is not that because it may be possible the number of companies which a company considering as bankrupt which is 32 in 2002. And the set of similar type of companies which are doing well which are non- bankrupt is 93. Similarly this 49; obviously, they are different from 32, then the choice set of non-bankrupt companies which is 98 is definitely different from 93.

They may be in the same sectors, but we will consider them as different; different such that the comparison which you are going to do with 32 with right 93 for 2002, 49 with 98 for 2003, 23 with 121 in 2004 are similar; that means, we can club good and bad companies depending on the fundamental concepts of the companies are similar. The reason why 93, 98, and 121 were taking in non-bankrupt companies was that; this was the closest based on which you can find the group such that, they can be matched with the bankrupt companies.

So, obviously there would be more companies which are non bankrupt, but the 93 98 and 121 are the closest to 32, 49, and 23. So obviously the below row gives the total which is 416 for the totals are number of companies, 312 and 104 for non-bankrupt and bankrupt companies.

Now, let us pause and think why have done this in order to separate bankrupt and non-bankrupt. Now remember one thing, that if I have been talking about for in the previous 15 minutes also and also I have mentioned time and again in the 55th class. That the concept on negative and positive selection were based on the fact, the negative one are

the ones this AIS concept are the companies or the so called entities I will use the word entities when I am discussing the concept of AIS in general are the entities which are as far as away from the concept based on which we are trying to match. And positive means they are as close. So, obviously when we build up the algorithm, it will be set of companies based on which we will do the clonal selection one set would be as bad as companies as possible and we replicate and increase it cloning clonal concept.

So, consider this 32 number of companies in 2002 were bad. So, we will basically expand or mutate and clone this 32 to larger numbers with different ratios such that the affinities remain the same. Such that we can match any other company which is closer to this 32 plus companies would definitely be matched as a sick company. Similarly 49 is cloned to a larger value; 23 is cloned to a larger value; to get the negative AIS and the negative clonal selection active or to activate it and run it.

Similarly when you are doing the positive selection, this 93, 98, and 121 would be taken as the sets in the three years respectively such the 93 would be replicated in the second year; 98 would be replicate; and in the third year 121 would be replicated to get the numbers or companies which are not sick such that the matching would be done. So, say for example, in the first set which and obviously, matching would be done; that means, those 93 has been mutated and expanded using the clonal positive clonal selection concept, 98 has been cloned and mutated and expanded depending on the positive selection.

Similarly 121 has been mutated and expanded based on the positive selection. Now when you consider these all things you will see it as you check the results. So, say for example, I pick up a company, which I want to check whether sick or not sick or there is same prediction. So, if I am running the negative algorithm or the negative clonal selection algorithm, that company would be taken with all its ratios matched with the 32 plus number.

Why I am using the word plus; because those 32 has been mutated to a larger sample. It is something to do with the concept of bootstrapping not exactly, but I have basically bootstrapping and increase it. So, that company whatever the new companies I select; I take that that string of variables or string of entities or characteristics which I need to match. So, they can be the ratios and all these things the market share, the sales value,

the profit, the loss whatever it is. So, if it is matched with this 32 plus if the matching is in that negative set is as far as possible.

Then obviously the company is good. So, we will throw it away; throw it means keep it aside and basically mark it as a good company. Now consider that same company I am trying to consider under the positive clonal selection; that means, 93 plus. That companies taken; obviously, I will take the same ratios as I have done for 32. Now the match will be the highest, because it falls under the category.

So, that will be not kept aside that will be considered in the set which is good. So, we will say again the company would definitely not be sick. So that means, we are doing both a negative comparison finding out the distance as far as possible keeping it aside saying that this is good one in the similar one when we do it the positive or positive clonal selection concept.

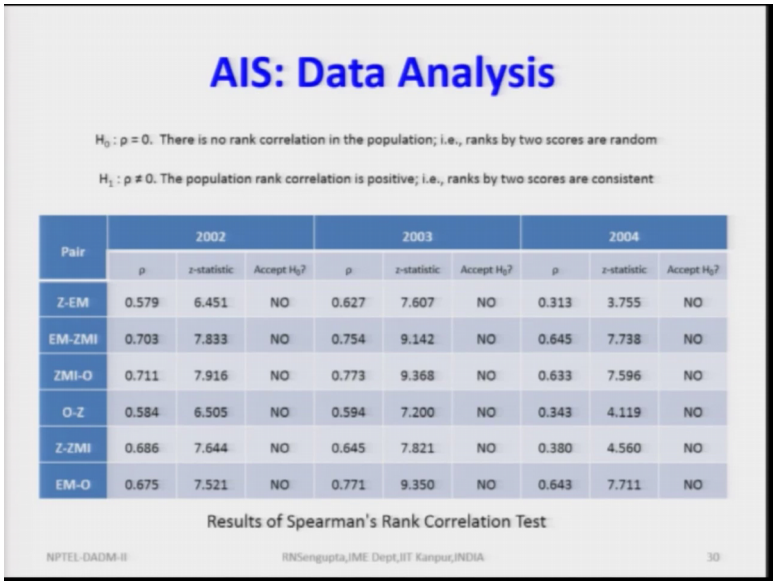
We find out the company is matching as close as possible again we keep in the in the other set which is good. Similarly if a negative, now why I am doing this or why we are doing in a negative matching and positive matching I am going to come to that within 2 minutes. Now consider company is bad and we are using the same set of 32 plus negative a bank bankrupt group and we are using the negative clonal selection a negative. So, the sick company comes it matches.

So, in that case it is kept as a in the set which is sick. When you take the same company and match it with 93 plus. So, the ratios and the values will be such that it will be farthest away it will be kept aside; that means, it is a bad company. So, in both way we are also trying to match it the good and the bad. Similarly we will do it for 49 plus, for 98 plus; plus means again I am repeating the clonal concept has being utilized to expand the set of comparison. 23 plus, 121 plus.

Now why I am doing this positive and negative comparison is that I want to basically overcome the errors. What are these errors I will discuss. I have discussed this type of errors in D A D M-I, but still I will basically try to reiterate with the diagram such that it is easy for us to consider this concept. This is something to do with the concept of hypothesis testing; because in hypothesis testing there would be some errors both positive and negative how I am going to utilize that. And I am trying to utilize the

concept of hypothesis testing, based on the concept of negative negative clonal and positive and positive clonal selection concept.

(Refer Slide Time: 24:26)




Now, here I will basically first try to utilize. So, this is a not very important for our; for the concept of AIS, but I will just give the results such that it was easy for comparison how good or bad the model is.

So, we will take the this compare the pair wise comparison we take the same set of companies. We use utilize those already developed model. The Zeta scores, the Zabrian score, the Olsen, score the Z score; so all these things are compared. So, this Z and EM is the Z score and the Emerging Market score. Then this O and Z is the Olsen score and the Z score. So, all these things are.

So, this the last one is E M is Emerging Market and the Olsen score. So, based on that we want to rank them with the ranking rank correlation as there; then how close or farther they are. So, we consider H naught where the rank where rho is 0 means there is no rank that is the rank by 2 scores are random; there is no match and if rho is not equal to 0 rank of 2 scores are consistent.

So, once we do that we find out this concept. So, we will be basically utilizing the concept to find out, whether we should consider H naught. So, H naught consideration being there; that means, there is some discrepancy in the comparison of this 2 different

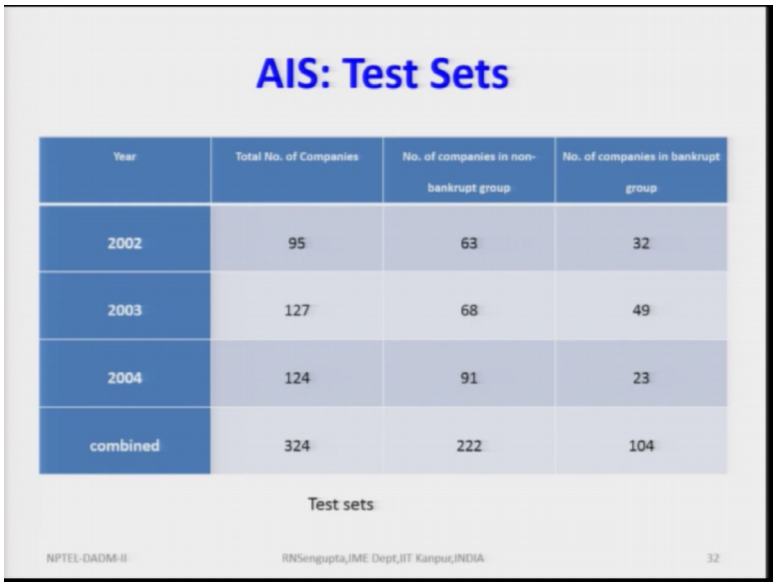
models when I am trying to basically find out any company which is bad. So, if they match in both the cases that mean yes it is sick under Z score also sick under E M algorithm also; that means, the company is really sick and the prediction of the model is good in both the cases. Very interestingly in any of the cases matching of these two algorithms everything is NO; that means, we do not accept H naught; that means, these ranks of these 2 scores or the comparison of a company by using 2 different scores are always same; that means, there is no discrepancy.

Which is good which means that underlying the models; obviously, they can give a good prediction we will check that. So, hence trying to compare a company which is sick, utilizing any 2 models gives us good robust result; that means, yes the company is really not performing well and it is sick. So, just; so, it can could have been done with any comparison with Z score or Emerging Marketing score, Olsen score could have been done with r model also; that means, the AIS and that matches that will match. Now we change this concept the difference between the so the predictive power.

So, we basically there is n the other case sign test or no difference test we do that again we find out that p. So, in that case matching and rho was is equal to zero and non zero. No, rho was equal to zero and non zero in this case we you use the case of sign test where p is 0.5 and p is not equal to 0.5.

So, if p is equal to 0.5 there is no difference between the 2 sets and we will accept the H naught; that means, here H naught has a opposite meaning with respect to H naught in the initial case. And p is equal to 0.5 means there is a difference not equal to 0.5 is there is a difference between the two sets of rank. And again very interestingly all the values are such that we will accept H naught as it should be; that means, using 2 different tests we are able to compare pair wise models as really being good in order to predict the sickness of the companies as it should be.

(Refer Slide Time: 28:28)




So, some of the combined values, number of companies, number of sick companies we take it and these are the test results. Once the test results are done, we will basically try to utilize in a AIS model. So, with this I will just close the 56th lecture and continue in 57th about the discussion what the results for the AIS are. Have a nice day and.

Thank you very much.