

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

Lecture – 57

AIS

A welcome back my dear friends and dear students a very good morning good afternoon good evening to all of you wherever you are. I am considering that some of you may be outside India taking this course; so that is why it can be at any point of time. Now this is the DADM-II which is Data Analysis and Decision Making-II lecture under the NPTEL MOOC.

And as you know that this total course is duration was for 12 weeks which is 60 lectures and which gets converted into 30 hours of lecture, 30 hours because each lecture is basically for half an hour and we are already completed 11 weeks. So, you already an each week with 5 lectures and you have already completed 11 assignments we are in the 12 th week and as you can see we are in the 57 th lecture.

So, with the end of this the set of lectures till 60 we will complete the; you will definitely complete the assignments till 12 and then we ready to take the question paper of the final examination. And my good name is Raghu Nandan Sengupta from IME Department IIT Kanpur in India. So, if you remember we are considering the sickness prediction considering the good and bad companies. And in the last class we discussed that how we can compare whether the prediction or the models themselves are good enough.

And we used two different tests rank tests and all the and such test to compare how good or bad the pair wise comparison of the results were and they came out to be very good for the set of companies which we took for 2002, 2003 and 2004. Now we basically set some parameters for now we are going to actually come to our AIS application. Now the AIS application will use the algorithm which you are decide not developed based on AIS concept we are basically made it fine tuned for our this application financial prediction.

So, the algorithm is negative along with the clonal both clubbed together positive and along with clonal they will be utilized. So obviously, if you remember for the clonal one we had the threshold r_1 , r_2 , r_3 then based on that we picked up from that NS non self one we picked up n_1 , n_2 , n_3 . So, we have to basically decide that.

(Refer Slide Time: 02:52)

AIS: Parameter Selection

Parameters	Procedure-I	Procedure-II	Procedure-III
n	100	300	900
r_1	47	47	47
r_2	30	30	25
n_1	N/A	N/A	10 ✓
n_2	N/A	N/A	250 ✓
ρ	N/A	N/A	5

Final Parameter Settings for Combined Data Analysis

Parameters	Procedure-I	Procedure-II	Procedure-III
n	300	900	900
r_1	47	47	47
r_2	30	30	30
n_1	N/A	N/A	30 ✓
n_2	N/A	N/A	750 ✓
ρ	N/A	N/A	5

Final Parameters selected for comparison with statistical models

NPTEL-DADIM-II RN/Sengupta, IIT Kanpur, INDIA 34

So, based on that we have in the procedures; procedures being the final setting of the combined data set and this combined data set can be of one set which matches one and another set which does not match we basically utilize that for both the sets. The parameters are we take in the final procedure III which is the combined one we take two threshold, this threshold need not be only two it can be more than that we take two thresholds r_1 and r_2 and the numbers are basically n_1 and n_2 which is basically the.

These are the parameters set for the combined data analysis and the final parameters based on which for the comparison for the statistical models utilized based on how we generate replicate again is given as n then we will take r_1 and r_2 and based on that we will proceed. Now remember one thing you may be thinking that here it is 300 while the values are 10 and 250 here the values 900 while the values are basically 30 and 750 which is not matching addition.

Yes, it may be possible because, if you take the threshold to n field r_3 level; obviously, some those threshold would be there. So, we are keeping some level of bound for the threshold such that there may be some elements or sets or individuals. In the I am considering the individuals of the companies which may not match so; obviously, those concept of threshold would not be applicable to them; is basically we have a sieve and we are basically straining them.

So, they may be some particles which basically go through the both the levels of strainer though the strain strains sieves r_c for example, of a value of r_1 and r_2 and r_1 and r_2 are such that they may be some particles which go through that we are not going to consider that. Then n_1 and n_2 are the number of particles which remain in the sieve accordingly.

(Refer Slide Time: 05:00)

AIS: General Hypothesis as Applicable for Problem

H_0 = All the companies detected by the detector set are bankrupt
 H_1 = No companies detected by the detector set are bankrupt

Type-I Error percentage = $\frac{\text{number of bankrupt companies not classified as bankrupt}}{\text{total number of bankrupt companies}} \times 100$

Type-II Error percentage = $\frac{\text{number of non-bankrupt companies classified as bankrupt}}{\text{total number of non-bankrupt companies}} \times 100$

NPTEL-DADA-II RN'Singupta, IIM Dept., IT Kanpur, INDIA 35

Now, once we compare, so, the comparison part I have already told again I repeat and why n field remember I said that we will do it comparison between both positive clonal negative clonal. So, again I will repeat a company is taken with all the ratios all the values of the financial front and we if we use the negative and we have already those 2002, 03, 04 good and bad sets non bankrupt good and bad. So, when we basically compare them using the negative or the positive a new company comes based on the ratios, the thresholds whatever decided we can change the thresholds, we are too strict or too relaxed would basically treated the value of the threshold.

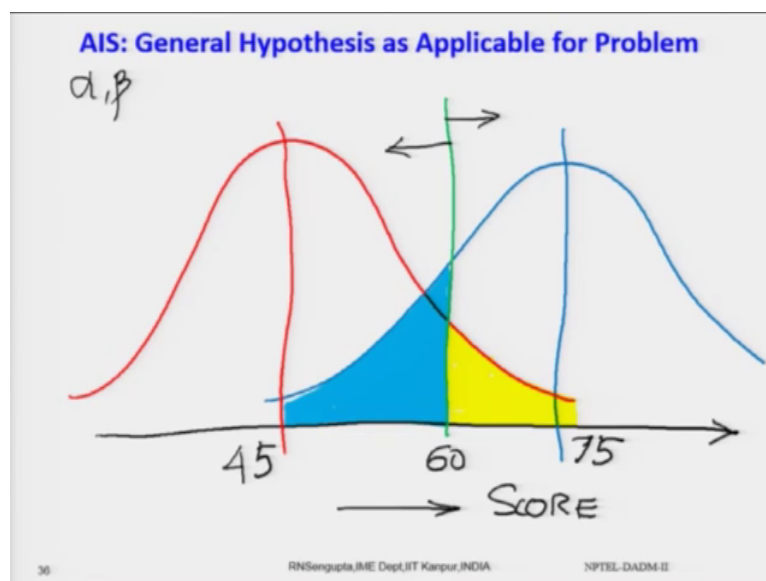
In the negative one if they do not match so, it is kept aside which means it is a good company and in the negative one if they match it is kept aside on the other side which is a bad company. Now this was the negative and negative clonal. Now whether the companies have good grouped properly will again I need for the positive one. So, the company which was set aside as a good company should now fall into the good set if you are using the positive selection criteria. And the company which was negative which was bad and under the

negative one would now basically be the complementary part; that means, it is known would be rejected by the good the positive selection algorithm.

Now, this if it matches; obviously, it will mean that there is no error in both these cases, but they would be errors; obviously, as we can see from a we will see from a diagram. So, the general hypothesis as applicable from the problems are H_0 is that all the companies directed by the detector set are bankrupt and H_1 is the number of companies detected by the detector set are no companies directed by the detector set are bankrupt. So, they will use to utilize this H_0 and H_1 accordingly H_0 means H_1 .

Now what are the errors would be a good company is analyzes bad and a bad company is analyses as good. So, it can be both the ways in both the algorithms. So, let me draw a diagram in order to clear this concept; let me know let me it is like this. So, this would go let me I have to basically draw it. So, the space should be there, create the slide properly. So, I can draw it easily and all of you can check yes done let me make the diagram ok.

(Refer Slide Time: 08:50)



Now, let us consider. So, I will first draw the I will give the background. So, consider I will give first take the example which I have already done in DADM-I and then cite a second example which is applicable here. And both are from the financial sector so, it will be easy for you to understand, very simple examples. Consider you are a bank manager you are a Chief Manager of a Bank in a Big City or in a Big Rural District or Rural Town and many loan applications come.

So, you have to basically take a decision whether you want to give take a decision to give the loan sanction loan or deny the decision. Now when you do that what you will do consider is that you will consider many of the parameters; that means, what is the age of the person what a business or company, where he is working ,whether he has a business whether he has own house or apartment land then whether he is paying in; obviously, they are paying income tax, but they will consider that what is his income based on and what is the tax is he or she is paying.

Whether he or she has his own vehicle, then whether the person has taken previous loans and what is the what type of sector the person is. So, all these things and what is the age and all these things some objective and some subjective criteria are considered. Based on that you basically set a score as 60; 60 score above you give the loan 60 and any score below 60 you deny the loan.

So, now graph would be like this, so this is the score line. So, they would be and we consider the that there are many people applying for the loan. So, the people who are in the long run would definitely pay the loan I will draw their distribution as normal with the color blue these are the people who will pay you back the loan. So, it is average score second set of people who even if you give the loan they would not return the loan, average is in the red color is red so; obviously, the score is less. Now consider that you have set for your initial score of 60 which I mentioned this is 60, so, this is the score.

So, I meant I am writing is as 60 and consider this score say for example 75 do not I am not drawn to scale consider this is 55 not 55 it is it should be. Now consider the problem I will use a different color to highlight it. Now consider this area this area let me mark as yellow with a highlighter. So, this portion yellow portion is for those set of people who if given a loan would not return and this set of people who even if they are good they are denied the loan.

So, they are denied the loan so; obviously, these would be two errors. So, the errors I will mention as alpha should not be such would alpha and beta is the errors. Now you will think that what if I decrease or increase 60? So, I take the line 60 on to the right or the left, but this is the if I taken to the right the blue portion increases the yellow portion decreases. Similarly, take a mark 60 on the left the blue portion decrease in the yellow portion increases in both the ways it is a loss for you for the bank manager why? Yellow means bad debt they would never

be returned, blue means opportunity cost opportunity loss for you business. So, both these are good customers gone what bad customers come in.

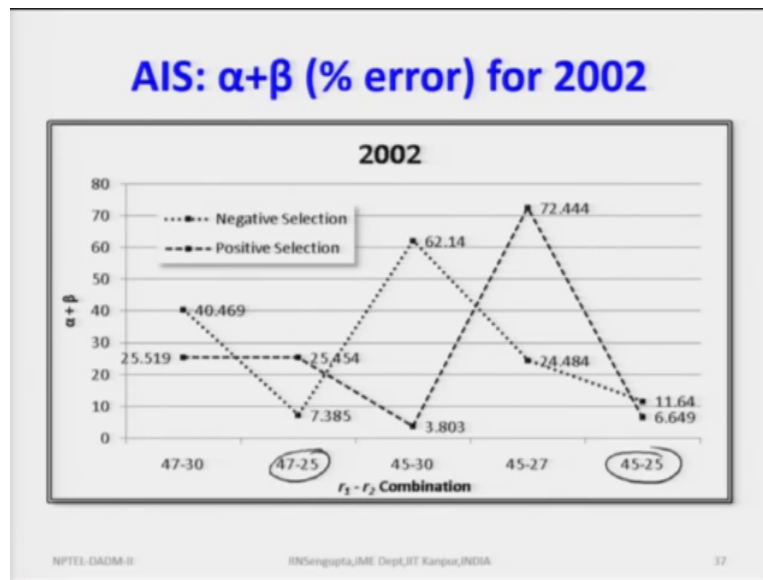
Now, trying to basically minimize a sum of alpha beta is the best way and we have methods how to tackle it in statistics hypothesis testing. Consider the same example now as I said I will consider the example that is applicable for the banks. So, banks consider the same type of distribution normal good companies bad companies you set a score threshold for yourself. Threshold means that the values which you have taken r_1, r_2, n_1, n_2 , so; obviously, you can make it more refine, but; obviously, you would not like to spend.

So, much of time in trying to do the refinement you want to double check how good or bad your results are. And if you remember the comparison which you did between two models taken two at a time and based on that you basically in one case you do not want to reject H_0 and in one case you want to reject H_0 and both the results in both the tests rank tests and all these things for selection test were coming fantastic.

So, in this case you will basically consider the companies which are bad and they have been predicted as bad in both the models negative selection positive selection and the companies which are good have been analyzed accordingly negative selection and positive selection. So, negative bad companies bankrupt companies prediction means in under the negative selection algorithm would be if they are bad they would be clubbed that is bad and when you take the positive selection they would not be clubbed in the good set, they would be kept aside.

And vice versa would be done for the case of the good companies if they are utilize in the negative selection algorithm they would be not clubbed as bankrupt they will kept aside and when using the positive selection algorithm they would be clubbed and kept in the set which are good companies. So, based on that if both of them match; obviously, the error is minimum.

(Refer Slide Time: 15:57)

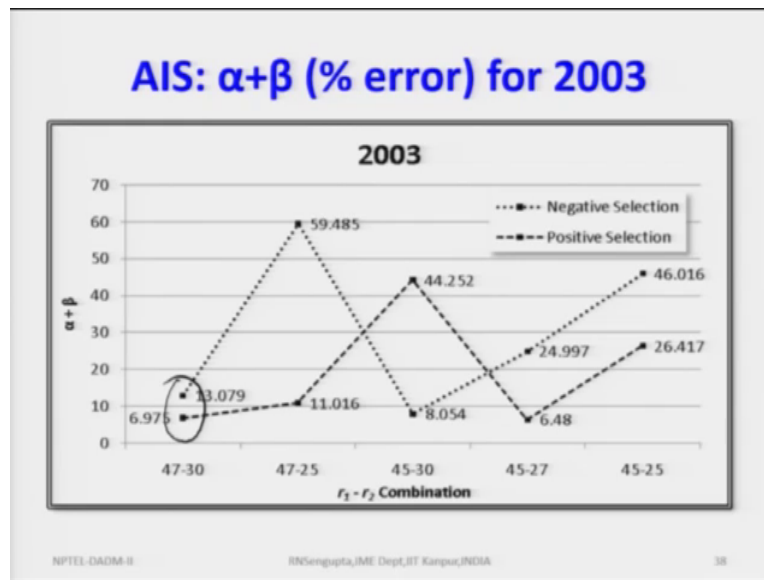


Now, we do a $r_1 r_2$ combination based on which we found out. So, r_1 and r_2 combination would be done in such a way. So, in the negative selection algorithm I am if you remember we are taking the sum of alpha plus beta. So, we are changing the ratios and if you consider the positive selection and negative selection algorithm for combinations of 45, 30; so these are the combination I am taking for $r_1, r_2, 47, 25, 45, 30, 45$. So, there can be different combinations, I am starting from 47-30 to 45-25.

It can be others also I can start in so; obviously, it is the groups would be 47 and then 3025, 3025, 3025 or 3040 so on and so forth. I am just taking arbitrarily few sets depending on the problem. So, I find out under the negative selection and the positive selection comparison the sum of alpha beta the errors.

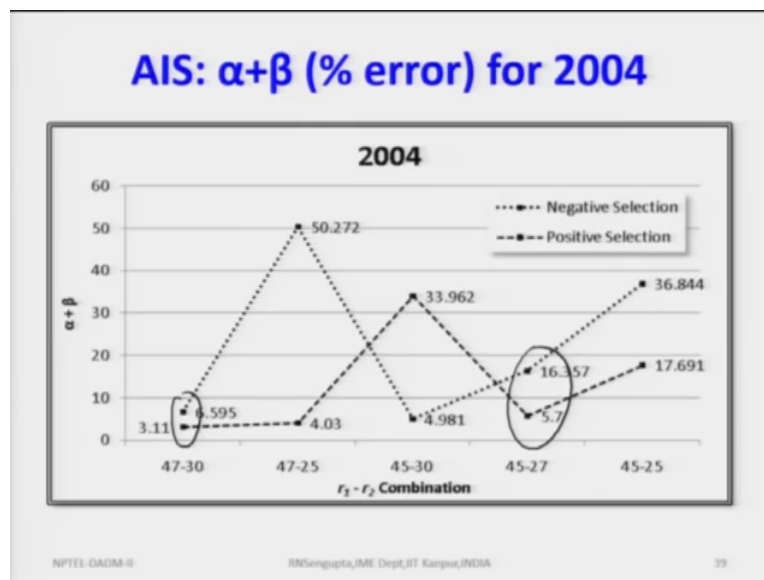
So, the errors in the percentage sense would be 40 I mean 40 percent, 25 percent, 25 percent, 7 percent, 62, 3, 72, 24, 11, 6. So, if you consider a comp a combination of values gives me a value of errors of 32 and 17. So, any combination better than that; obviously, would be great. So, I do it for 2002.

(Refer Slide Time: 17:23)



Similarly, I will do it for 2003. So, the combinations are this is the best one 19 other values are quite high.

(Refer Slide Time: 17:37)



Similarly, I do it for 2004 values are 9 this value is also decent enough; obviously, there would be an error these are errors because you can definitely fine tune the model depending on your r_1, r_2 combinations and the collection of sickness.

So, may be possible some the companies we are taking as sick or good companies may not actually be that so; obviously, in the data set based on which you are trying to proceed there would be errors. So, we take the errors for type I and type II percentage wise.

(Refer Slide Time: 18:13)

AIS: Combined Data Analysis

Procedure	Type-I Error (%)		Type-II Error (%)	
	average	St. error	average	St. error
Procedure-I	1.943	0.017	1.572	0.028
Procedure-II	1.954	0.016	0.984	0.011
Procedure-III				

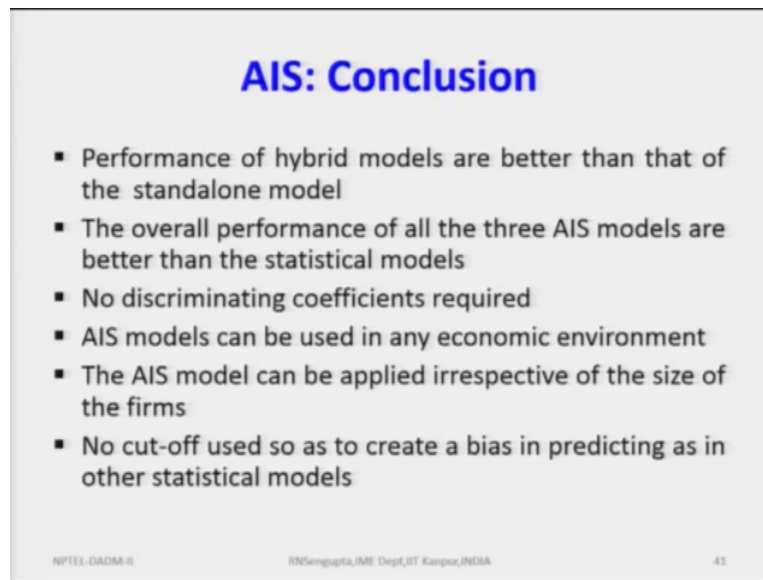
Classification results for combined data with $r_1=47$ and $r_2=30$

NPTEL-DADM-II IITKangra, IIT Roorkee, INDIA 40

Under procedure I and procedure II the companies when we do the averages of the type I and type II errors are coming out to be about 2 percent and type II errors is about 1.5 and 1 is 0.9 I am considering as 1 and the standard would also give you amount of dispersion which you have.

So, this is the classified results for 3047 and 30 the compare comparison which we got the best results. Now why you are seeing that why do get an average and the standard data? With those value r_1 and r_2 , I repeat the algorithm. So, is basic random generation number. So, clone clones are being generated randomly or the mutations are happening randomly. So, based on that it gives me the average values; so I get decent results based on the values.

(Refer Slide Time: 19:15)



AIS: Conclusion

- Performance of hybrid models are better than that of the standalone model
- The overall performance of all the three AIS models are better than the statistical models
- No discriminating coefficients required
- AIS models can be used in any economic environment
- The AIS model can be applied irrespective of the size of the firms
- No cut-off used so as to create a bias in predicting as in other statistical models

NPTEL-DADIM-II RNSengupta, IIT Kanpur, IIT Kanpur, IIT Kanpur 41

So, to conclude the AIS conclusions or the Artificial Immune System conclusions are performance of the hybrid models are better than that of the standard model. So, one can utilize a hybrid one I am not given the detailed results it means basically it would be a combination of clonal one before that we do the negative and the positive one at a time or we use the positive and the negative one at a time. So, we go first negative you select the companies as bad.

So, they are grouped in the s set another's are kept in the complementary set of s which are good then again we pass the same things under the positive selection one. So, the set which is in complementary one they would be put in now the t set; t which is the good set of good ones which we keep and initially s and s dash t and t dash would be empty. Similarly, the s which were bad when they go into the positive selection they are put in the t dash set; so if you utilize that the results are much better.

The overall performance and all the three AIS models are better than the statistical models those results which gave. No discriminating coefficients were required. So, they values based on which we to got the results were good. AIS model can be used in any economic activities or other type of problems they can be applied irrespective of the size of the forms. So, size of the form did not matter here, we checked it and there are subjective and objective criteria's also which can be done. No cut off was used to create a bias in predicting as in statistical model which values.

So, with this I will conclude the; this 2nd lecture for the last week and try to come and go into the other topics accordingly for the next three lecture, have a nice day.

And thank you very much.