

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

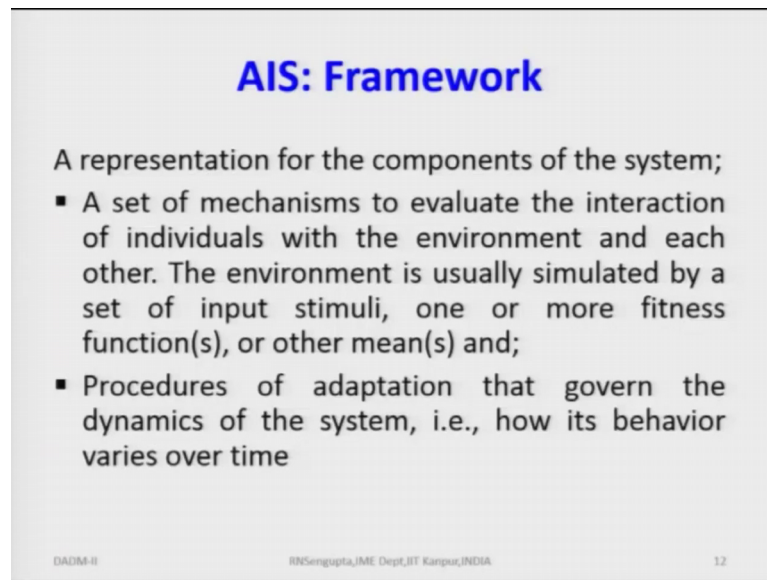
**Lecture – 55**  
**AIS**

Welcome back my dear friends very good morning, good afternoon and good evening to all of you where will you are in this part of the world. And as you know this is the DADM-II which is Data Analysis and Decision Making-II course under the NPTEL MOOC series. And we are in the last class for the 11th week and this total course duration as you know which is still before I started the lecture is for 12 weeks which is actually of 30 contact hours which gets converted into 60 lectures because each lecture is for half an hour.

And each week we have 5 lectures and after each week we have assignment. So, with this end of this lecture you will have the 11th week assignment and then next week you have the 12th week the last course for the wrapping up. So, if we and my good name is Raghu Nandan Sengupta from IME Department at the IIT Kanpur. So, if you remember we discussing about the concept of Artificial Immune System the concept of positive selection, negative selection and how the general concept of antibodies, antigens and pathogens.

Basically are the building block the ideas of the building block based on which the artificial immune system can be build up. And we will further discuss that with an example either today or in the next class which will be in the first lecture of the 12th week.

(Refer Slide Time: 01:38)



**AIS: Framework**

A representation for the components of the system;

- A set of mechanisms to evaluate the interaction of individuals with the environment and each other. The environment is usually simulated by a set of input stimuli, one or more fitness function(s), or other mean(s) and;
- Procedures of adaptation that govern the dynamics of the system, i.e., how its behavior varies over time

DAADM-II RNSingupta,JME Dept,JIT Kanpur,INDIA 12

So, for the general framework for the AIS which is the Artificial Immune System is a representation for the components of the system is very essential which means a set of mechanism is to be evaluated such that the interaction the individuals with the environment and each other can be made.

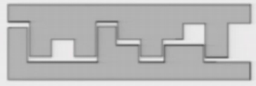
So, if you remember they can be a primary and secondary reactions or the fighting mechanism in the primary one pathogens come and the antigens or the good sales combine them. And in the secondary one depending on some priority information set the cells are ready to fight the invading virus or the bacteria. The environment is usually simulated by set up input stimuli can be either high temperature, low temperature, cough and cold so on and so, based on which the stimulus occurs for the replication of the cells.

In set of input stimuli one or more fitness functions could be utilized in place of the stimuli in the actual utilization or other means can be also utilized. A procedure for the adaptation that govern the dynamics of the system that is how its behavior will vary or evolve with time would also be briefly discussed and analyzed.

(Refer Slide Time: 02:56)

**AIS: Idea and Representation**

- Representations
  - ❖  $m = \langle m_1, m_2, \dots, m_L \rangle$ ,
  - ❖  $m \in S_L$
- Describe interactions between molecules
- Degree of binding (Affinity) between molecules
- Distance measures
- Complement threshold



DADM-II RNSengupta, JME Dept, IIT Kanpur, INDIA 13

Now, whenever you are trying to basically find out the best free state function and how the replication can be done or how good or bad the solutions are with respect do what you want to achieve whether maximize, minimize or want to basically have some combination of them.

We will basically have different concepts of distance function. So, distance function can be quadratic, can be 1 1 norm, can be 1 3 norm, can be infinite norm, can be (Refer Time: 03:21) distance, can be in from the point of view of statistics, can be hamming distance. So, all this things can be compared. We for our problem in this case when you are trying to solve the problem we will consider the hamming distance for our discussion.

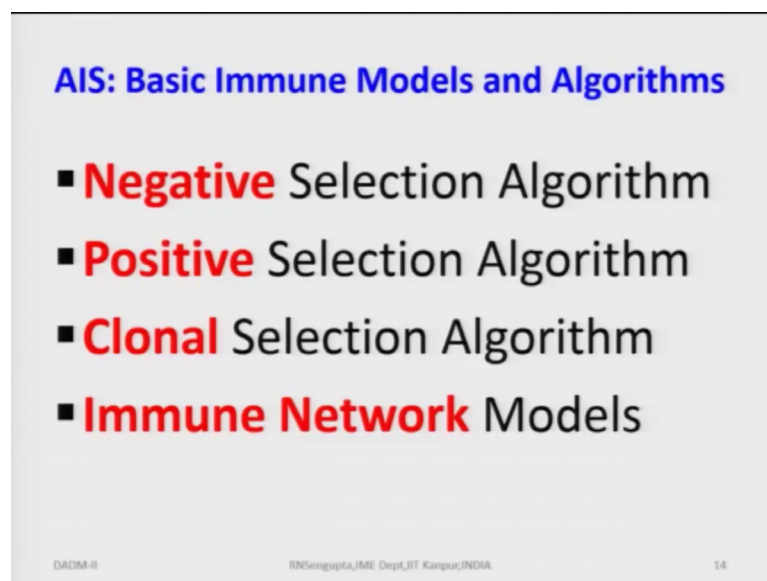
So, representation would be done by a vector of size of say for example, 1; 1 would be all the individual component based on which the characteristics of the representation of the similarity dissimilarity can be done. So, it will describe the interaction between the molecules; molecules means the like say for example, there are attacks happening from three different types of viruses. And we will consider each virus has some characteristics based on which the fighting WBCs would be build up and consider the distance function or fitness function based on which we will try to build up the model is four in number. So, the first virus that would have four such matching to be done.

Similarly, for the second one and the third one; so it describes the interaction between the molecules and the degree of binding on the affinity of between the molecules is also be

based on this fitness function. For the fitness function if you remember I would consider the concept of distance measures and I did mention what are the distance measures like the  $l_1$  norm,  $l_2$  norm,  $l_3$  norm,  $l_\infty$  norm and all these things. And we will basically have a threshold based on which we will try to find out how close or far the fitness function is it need not be the fitness function fit is the best they may be error like epsilon errors.

But what is the value of epsilon based on which we will take a decision, yes this is a good fitness function and we should proceed or no it is not a good fitness function based on the tolerance limit; we should not proceed we will basically we should basically be able to say that before and before we analyze the problem and based on that we build our methodology of AIS.

(Refer Slide Time: 05:14)



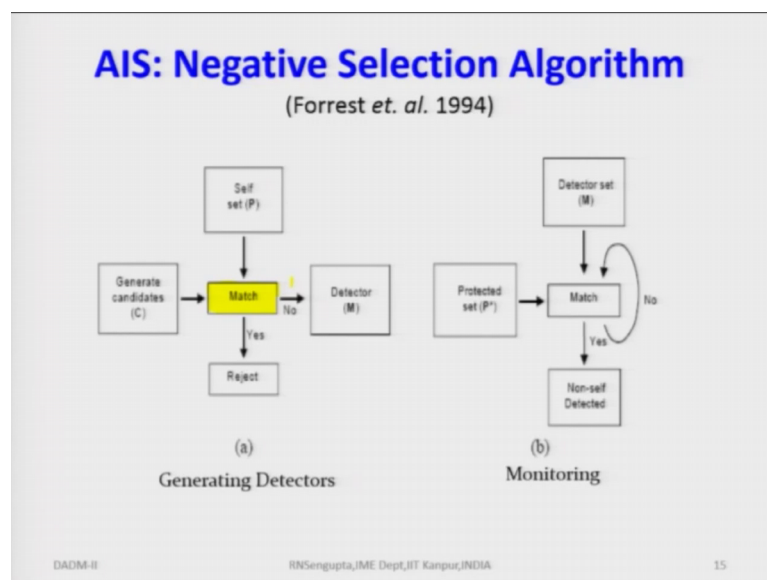
So, generally the basic immunes models and the algorithms are based on the negative selection algorithm, positive selection algorithm, what are this if you remember in the last class in the 54th class I did mentioned.

That in the positive and negative we start with the bad set and build up the on that and in the positive in the negative selection we have start with the bad set in build up from that and in the positive selection we will build start with the good set and build up on that. Clonal selection would also be considered where the replication of the cells or replication

of the so called algorithms or the optimization problem would be build up on a set of say for example, a seed.

Seed means, the when you start of a random number generation or say for example, you want to generate some numbers you would basically start with the seat of the value based on which will build up those values. So, based on a seed you will basically try to replicate how good or bad cells can be build up such that your fighting mechanism can be build up in this case the fighting mechanism is basically how close or far it is from the fitness function. And the immune network models would also be build up based on this a negative selection, positive selection and clonal selection algorithms.

(Refer Slide Time: 06:26)



Now, this negative selection problem we will take the general formula idea from forrest us forrests paper in 1994. So, the process I will just give you the general process how it works. So, the process basically works on the idea that you will basically have a matching set I will describe the matching set later on, but. So, you will basically have a matching set. So, the what is the matching set I am going to come to that later and some generated candidates are coming; that means, generated candidates means I have cell which you wants to fight and then the attack is happening from the external sources.

So, this we will consider the as the generated candidates who come. I match I find out the matching is good, in that case virus is a coming my antibodies are there, my antigens

are there, antigens fight they find the matching is good they club together deactivate each other and they become dead.

That is process that is what it was the general idea. So, in this case once the generated candidates comes and what you do is basically you consider a self; self means the type of WBCs which is going to fight. So, in this case what you will do is that you will basically have a so called generated initial solution. So, that would basically will be the self and the candidates who are coming are basically the different type of feasible regions of the solution space which will consider that whether they are good or bad.

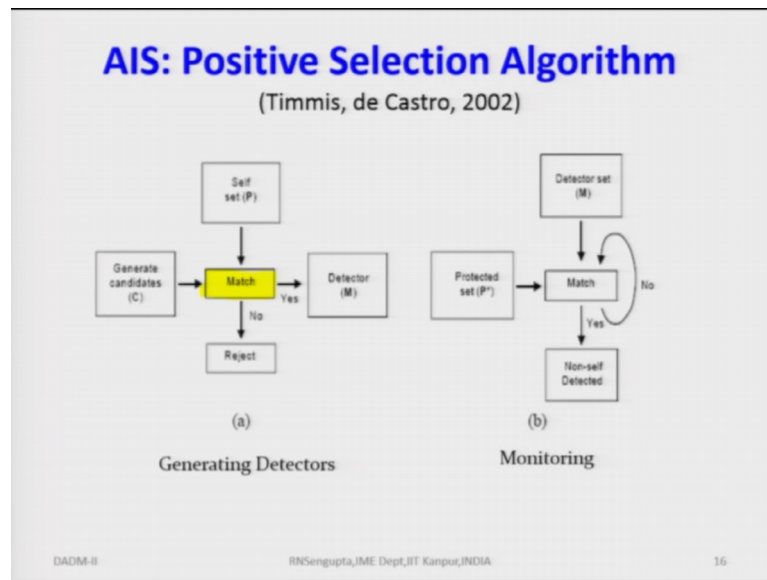
So, you will basically consider a self type of solution which you think is good that is good or not that is a different question you think that is a good based on the prior a set of informations. And you basically try to compare that the solution which you had with the generated solution which are coming from the feasible set. If they match you throw them away, if they do not match you basically put them in a separate set which is basically denoted by  $s$ . So, initially  $s$  is an empty set, null set and you try to basically keep those elements in this set.

So, what is happening is in general, negative situation means that I will throw away the so called good ones and keep the bad ones. So, in the process as I do the solution which are thrown away; obviously, would be the set of solutions amongst to each I will get the best solution. So, this is basically I am trying to go in a negative direction negative direction means when I am trying to do away with the solutions and basically choose one among those. So, once this set has been developed the detector set has been developed we will basically match those detector set with the existing solutions and so detector sets are the negative ones remember that.

So, you will basically match the competence level or the closeness level of the detector ones and obviously, it will be as far as possible based on that we will try to replicate and get how such solutions can be detected. And made more in number such that the solutions which are being rejected in the negative solution algorithms would slowly build up; such that we are able to use the best amongst the negative set of solutions which is the so called best optimum solutions.

Now in the positive selection procedure the method is just happens in the other way. So, how it is does it done.

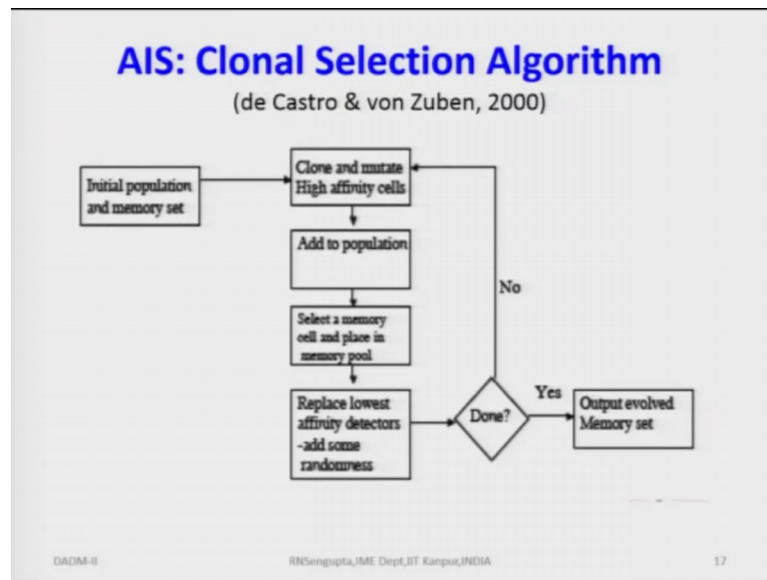
(Refer Slide Time: 10:00)



So, we basically take a matching set against the sets of solution generated candidates which are coming are mapped. But now if they match we put them in that asset initially we were putting only the bad elements in the asset; now we will put the good elements in the asset and throw away the bad ones. And based on the detector set which has been basically which was  $s$  which builds up with the good values.

We now basically replicate the problem and exactly the same way as we have done in the previous case for the negative selection algorithm. So, in the negative selection algorithm, we keep in the detector set the bad solutions and in the positive selection algorithm, we keep in the set as the good solutions and based on that we replicate and find try to find out the solutions. So, I will come to the example in few minutes.

(Refer Slide Time: 10:57)



Now, in the clonal selection algorithm the process will build up. So, how does the process build up? So, this is if you remember the clones would be build up or they would be generate in such a way that in the initial problem in the actual medical concept; we said that they should build up such that they are able to fight the pathogens which will be coming to attack the system or attack the body.

So, we will try to basically replicat the same replicate the same concept in order to build a solution set in a much better way. So, the initial populations of the memory set or the set based on which will try to replicate would be considered and we basically clone them; that means, we will have a in infinite sets of such near optimum not near optimum solutions near feasible solutions all feasible, but they are basically compute so through each other.

They would be build up and added to the population. So, once you select memory cell and place in the memory pool. So, say for example, you increase the population increasing the population means that you are trying to replicate it as close as possible to the probability distribution function of the population or the probability mass function of the population.

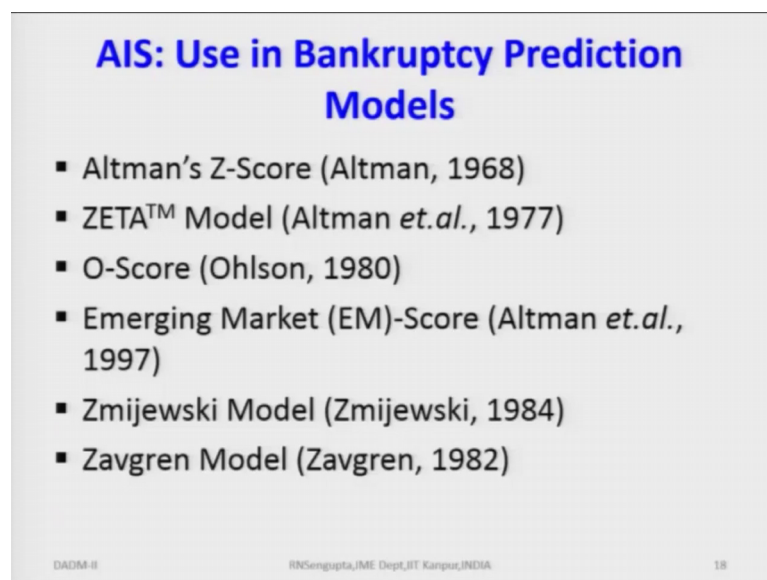
And based on that you will basically you will you replace as your replacing them your replacing the less affinity once when; that means, the ones which have the least matching. So, as you do that you basically build up population which of the almost the



same characteristics based on which you can replicate and find out the best solution in the best possible manner; that means, you are slowly building up your population.

Doing away with the ones which are further away from your actual population and basically build up a solution which are near optimum; optimal and based on that you can build up your model. Now we will basically try to consider in our case problem in the area of finance; so to give a brief background. So, we will use the concept of a negative selection algorithm, positive selection algorithm, clonal algorithm, used the concept of antigens, antibodies, pathogens and how they can be utilized in trying to basically build up a very simple detecting model in the area of finance.

(Refer Slide Time: 13:12)



**AIS: Use in Bankruptcy Prediction Models**

- Altman's Z-Score (Altman, 1968)
- ZETA™ Model (Altman *et.al.*, 1977)
- O-Score (Ohlson, 1980)
- Emerging Market (EM)-Score (Altman *et.al.*, 1997)
- Zmijewski Model (Zmijewski, 1984)
- Zavgren Model (Zavgren, 1982)

DADM-II RNSengupta,JME Dept,JIT Kanpur,INDIA 18

So, in generally Bankruptcy Prediction Models we have different type of models which have been developed starting the 1960s the main person who are the four the frontier person in the area of building of such model was by the name of Altman.

So, Altman basically propose the Z-Score in the 1960s and they were other scores also developed by Altman and is set of researchers one was the Zeta model which is now used as a trademark by many of the banks. Then you have the Ohlson Score which is the O-Score then you have again another Altman's model which is the Emerging Market model where the models for more specifically in emerging market as the economy developed.

We have the Zavgren model, Zmijewski models and all these things are part and parcel of the bankruptcy predictions model and they are all mathematically in nature and starting from 1960s to 1980s and 1990s, the models were developed in a big way. So, these are all statistical methods in general.

(Refer Slide Time: 14:13)

**AIS: Main idea is to Distinguish between Self v/s Non-self**

- A problem of *distinguishing between the **self** and the **non-self***
- A problem of *distinguishing between the **sick** and the **non-sick** companies*
- The Metaphor
  - ❖ Sick companies can be considered as the **antigens**
  - ❖ Non-sick can be considered as **self-cell**

DADM-II RNSengupta, IIM Dept, IIT Kanpur, INDIA 19

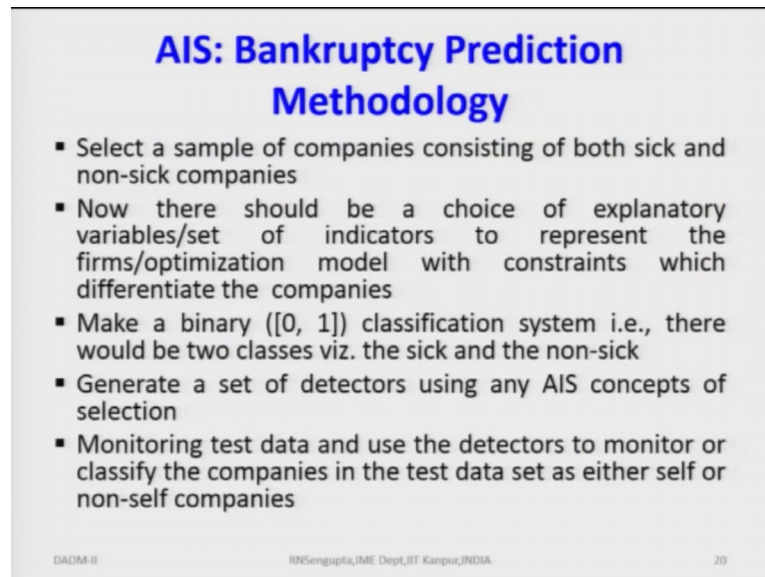
Now, we want to basically utilize the concept of AIS to build up a prediction model of bankruptcy and compare our model with respect to any of the existing models like the Altman or the Zeta model or the Ohlson Scores or the emerging market scores and all these things.

And see how good or bad our models are with respect to the general sample we are going to take to predict this bankruptcy. Now here what is the main idea for our problem is to basically the problem is basically to distinguish between the self and the non-self; that means, the good and the bad companies or the bad or the good companies depending on how you have been able to define the concept of negative selection and positive selection. So, the problem would be to basically detect the sick companies from the non-sick companies or the non-sick companies from the sick companies depending on how you as I am just mentioned few seconds back.

You have which model you are going to use in order to build up the detection algorithm. So, the metaphor for our example would be sick companies can be considered as the antigens and the non-sick can be considered the cells which will basically replica

themselves in order to basically give a much better prediction how the good companies are doing based on many of the parameters so; obviously, the question will come that what are the parameters we will consider to build up on model.

(Refer Slide Time: 15:34)



**AIS: Bankruptcy Prediction Methodology**

- Select a sample of companies consisting of both sick and non-sick companies
- Now there should be a choice of explanatory variables/set of indicators to represent the firms/optimization model with constraints which differentiate the companies
- Make a binary  $[0, 1]$  classification system i.e., there would be two classes viz. the sick and the non-sick
- Generate a set of detectors using any AIS concepts of selection
- Monitoring test data and use the detectors to monitor or classify the companies in the test data set as either self or non-self companies

DAOM-II RNSengupta, IIM Dept, IIT Kanpur, INDIA 20

So, in the Bankruptcy Prediction Methodology, the concept would go alike this we select a sample or companies consisting of both sick and non sick companies; so it will basically be both. Now there should be a choice of explanatory variables or the set of indicators which will be utilize to represent how good the bad the company is with respect to is whether is going to fail or not fail.

So, the parameters can be the assets, can be the market share, can be the balance sheet variables, can be the depreciation value, can be the amount of loan the company has taken, can be the stock market, can be the land prices, can be the ratio of the sales to total assets there can be different things.

So, in general in accounting we have about 14 ratios based on which we can analyze how the company is you doing good or bad. Apart from that as I mentioned the market share concept the concept of stock prices all this can be utilized to predict how good or bad the company is doing in the coming days and how it will do in the future. Now there should be choice of a as I said there should be choice of explanatory variables, sets of indicators to represent the forms, optimization model with constraints, which differentiate the

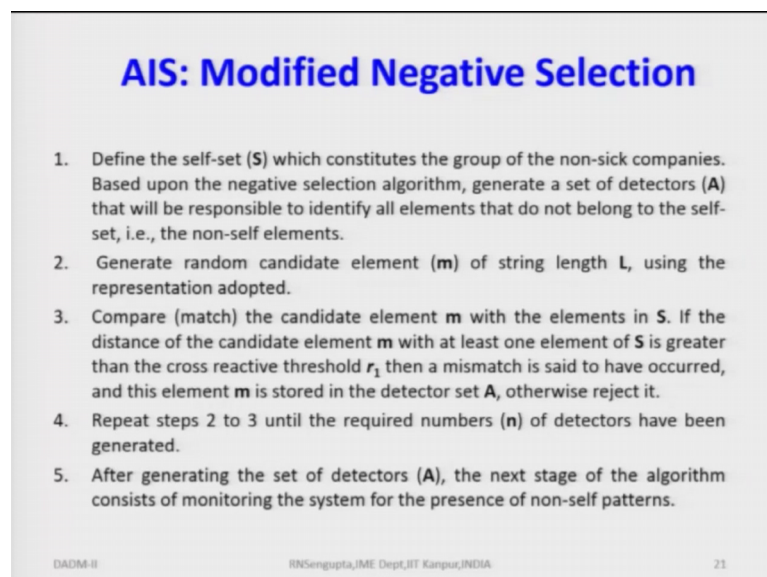
companies which are good from the bad and these variable should be such that they would be able to give us a much better idea.

So, what we will do is that we will basically make a binary classification system. So, that would be there would be two classes. So, one of the classes would be sick another classes would be non sick companies point 1, Point number 2 is that say for example, all these 14 15 this would be consider we will basically have a distance measure and try to find out how good or bad this distance measures are when we basically compare one of those parameters or any one of the characteristics based on which we can analyze whether the companies doing or good or bad.

So, I am not saying that there would be only one characteristics, there would be 14 to 15 different ratios we will try to compare these ratios individual and then try to basically find out the best combination which will give the set of class which are good and the set of class which are not good; that means, sick and not sick. We will generate a set of detectors using the AIS concept of selection as I mentioned.

So, those mapping would be done for the ratios on to the AIS measuring scheme or the distance scheme. We will monitor the test data and use the detectors to monitor or classify the companies into the of the companies into the test data set depending on the fact whether we are going to take the negative selection algorithm or the positive selection algorithm.

(Refer Slide Time: 18:08)



**AIS: Modified Negative Selection**

1. Define the self-set (**S**) which constitutes the group of the non-sick companies. Based upon the negative selection algorithm, generate a set of detectors (**A**) that will be responsible to identify all elements that do not belong to the self-set, i.e., the non-self elements.
2. Generate random candidate element (**m**) of string length **L**, using the representation adopted.
3. Compare (match) the candidate element **m** with the elements in **S**. If the distance of the candidate element **m** with at least one element of **S** is greater than the cross reactive threshold  $r_1$  then a mismatch is said to have occurred, and this element **m** is stored in the detector set **A**, otherwise reject it.
4. Repeat steps 2 to 3 until the required numbers (**n**) of detectors have been generated.
5. After generating the set of detectors (**A**), the next stage of the algorithm consists of monitoring the system for the presence of non-self patterns.

DADM-II RNSengupta, IIT Kanpur, INDIA 21

In the negative selection algorithm, the process we will work like this. So, we will define the set as which constitutes the group or non sick companies; that means, who are good. Based upon the negative selection algorithm generate a set of detectors A. So, the detectors means what are the variables based on which you will study. So, these will be responsible to identify all the elements that do not belong to the self set or the non self elements which are there. You will generate random elements m of the string of length L using the representation adopted you will compare and match the candidates element m with the elements in S.

They compare them how good or bad there if the distance of the candidates elements m with at least one element in S is greater than the crossover or the threshold value  $r$ . So, it can be  $r_1, r_2, r_3, r_4$ , depending on how you have been able to represent that, then a mismatch would be said to be have to have occurred and this element m is stored in the detector set A otherwise it will be rejected. So, this is the negative one in the other case positive one it will be just be the opposite.

You will repeat step 2 to 3 until the required number n of detectors have been generated and after generation of the set of detectors A, the next set of algorithm consist would be consisting of monitoring the system for the presence of non-self pattern such that how good or bad the repetition would be in the future.

(Refer Slide Time: 19:29)

## AIS: Modified Negative Selection

```
Algorithm: Modified Negative Selection
Input: self set S, cross reactivity threshold r, no. of detectors required n, and length of string L
Output: Detector Set A
Begin
  j ← 0
  While j ≤ n do
    m ← random()
    for each s of S do
      distance ← match(m, s)
      If distance ≥ r then
        Insert m in A
        breakFor
      endif
    endFor
    j ← j+1
  endwhile
  return A
end
```

DADM-II RNSengupta, JME Dept, IIT Kanpur, INDIA 22

So, this is the actual negative selection algorithm which will be utilized. So, the input set would be the set of  $S$  consists of reactivity threshold or the number of detectors required which would be of length  $L$  and the length of the string based on which will do the comparison  $L$  in number. So, these are all numeric values.

So, the output what you want to do is that basically have a non empty set  $A$  which would basically consist of the detector set. So, generally we will take the number of detector starting from 0. So, you put  $j$  as 0 and continue to doing that for each and every comparison till  $j$  is  $n$ . Now we will consider that  $m$  as the random variables and we will basically compare the distance function between  $m$  and  $s$ . So,  $m$  is basically that the set where we will try to basically put this detectors and non detectors depending on the negative the positive selection and  $s$  would be the comparison set and if that and you; obviously, have a distance function.

So, if  $m$  and  $s$  elements have a distance function greater than less than you will basically put those values from the  $m$  into the  $s$  set or the or outside the  $s$  set depending on whether the matching is less than equal to sum epsilon value for the distance or greater than that. So, you will continue that then you will basically increase the  $j$  to  $j$  is equal to 0 to 1 again compare for all the steps and continue doing that till you exhaust all the values of  $j$ . Now we will basically find out the modified negative selection followed by the positive selection model.

(Refer Slide Time: 21:07)

### AIS: Modified Negative Selection followed by Positive Selection

1. The non-self set (**NS**), which constitutes the group of the sick companies, is used to extract a refined detector set (**A**) from the set of detectors (**M**) that is generated by the modified negative selection algorithm. This enables the elements of the set (**A**) to additionally be able to identify at least one non-self element or a sick company.
2. Compare (match) the each element **m** of **M** with the elements in **NS**. If the distance of the candidate element **m** with at least one element of **NS** is less than the cross reactive threshold of  $r_2$  then a match is said to have occurred, and this element **m** is stored in the detector set **A**, otherwise remove this element.
3. Repeat step 2 until each and every detector in the set **M** has been checked.
4. After generating the set of detectors (**A**), the next stage of the algorithm consists of monitoring the system for the presence of non-self patterns.

DAAD-ii RNSengupta, IIT Kanpur, INDIA 23

So, what you are trying to do is basically your trying to replicate that first you do the negative one then you do the positive one. Now there is a reason for that, reason being that I will just talk about here before I proceed. Say for example, I am trying to detect company which is good. So, I do not know the actual whether the company is good or bad, but I think based on that variables I want to basically pass some judgment.

Now it may be possible that company even though it being good it can be clubbed as a bad set; that means, there is an error, on the other hand a company is bad and I know the and technically it should be bad such that there is no error, but it may be possible that the bad companies clubbed as good.

So, in this both the cases we have a type of errors which are known as the alpha and beta and which is basically I am taking from the statistical point of view. The general concepts it would be they are trying to basically minimize both alpha and beta separately is not possible. So, what will try to do is basically try to analyze the error such that the sum of the errors of alpha plus beta would be minimum.

So, when you are trying to do that you will basically have the type-I and type-II error under the concept of  $h$  naught on  $h_a$  which we have already done and in DADM-I. So, once that value of alpha and beta is given we will try to minimize that. So, generally the concept is done in such a way that the minimization is done keeping beta fixed we will basically try to minimize the values of alpha as far as possible.

So, this is the general concept which we will try to for basically follow for our case. Now in the positive selection algorithm what as the thing which you have done for the negative selection is basically do the negative way or the distance function be match such that we put away which are not matching in the other set. But in the positive set we will try to basically follow that we will club those which follow as soon as close as possible we will put in that set and require and then basically try to match them as close as possible.

So, in the general formulation which we are going to do is that we will first run the negative selection algorithm then the positive selection algorithm and in the other way we will basically run the positive selection algorithm then run the negative selection algorithm. So, with this I will close this 11th week class and start discussing about the problem in the first class of the last week and hopefully with the end of AIS which would

happen in the 56 class, we will be able to cover another one or one and half topics in the area of metaheuristic techniques have a nice day.

And thank you very much.