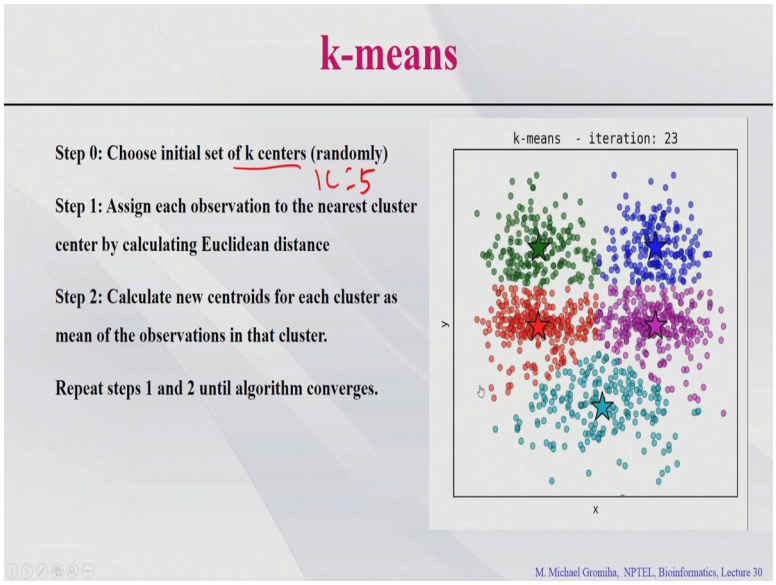
Bioinformatics
Prof. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

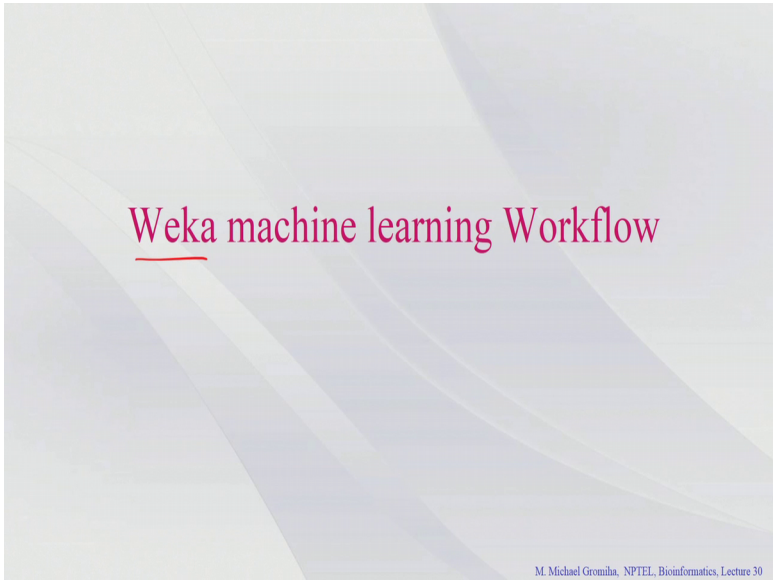
Lecture – 30b
Applications of bioinformatics II

(Refer Slide Time: 00:16)



So, now there are many machine techniques available. So, it is very difficult to write a code for each one of them and then check the performance. So, there is a common platform called the Weka.

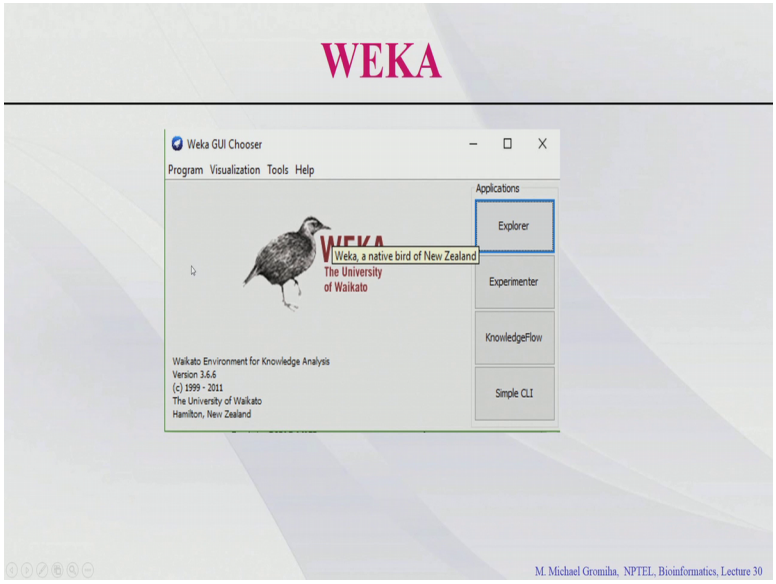
(Refer Slide Time: 00:29)



So, it is a platform which includes several machine learning techniques; currently it has more than 50 machine learning techniques including neural networks, support vector machines and clustering; Various methods are available in this weka platform. So, you can use any of these techniques and it easy to use it because you can choose it and you can see the performance.

Depending upon your problem, it will give you the different types of performance; I will explain how we use the weka machine learning workflow to classify these two types of proteins. For example transmembrane helical proteins and the transmembrane strand proteins.

(Refer Slide Time: 01:05)

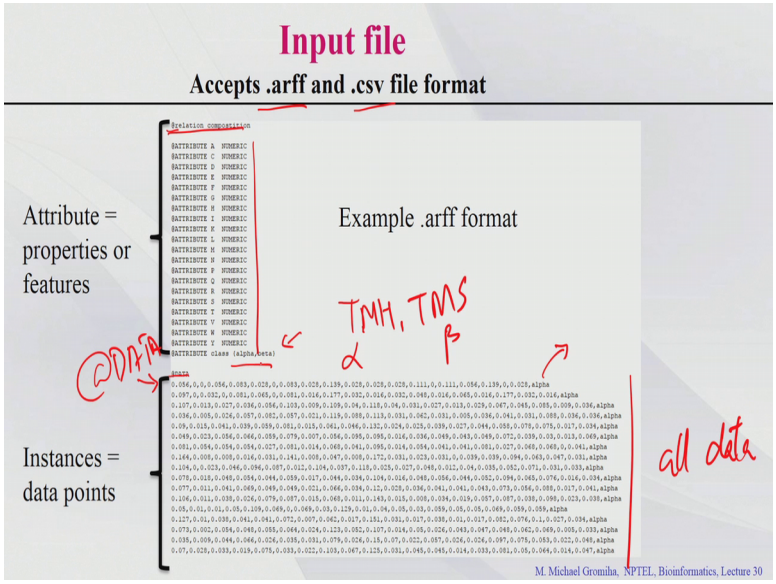


So, it has developed in the University of Waikato; so that is in a New Zealand. So, that developed long time ago, currently included various new features. When they originally initiated this weka they have only few features and feature selection algorithm was not implemented and few machine learning techniques; now they try to use it for some similar set of proteins.

Now, different instances and different sets of problems they incorporated various algorithms plus the feature selection, the attributes. So, now it is having a full fletched platform for any classification as well as the real value prediction algorithms. So, they have different applications or we can see the explorer or experimenter knowledge flow and the simple CLI.

So, if you go with the explorer then we can open a window; there you can manipulate your data give the input data and you can play around this weka with the different machine learning techniques; as well as your features, feature selection and the performance.

(Refer Slide Time: 02:07)



Then first I will tell you how to prepare the input file; early in the previous version of the weka, they accepted only the arff format and currently they also use CSV file format; you can directly get from the excel. Early days if you have to prepare the input in such a way that there should be in a particular format type that is arff format. If you see the arff format; first we give the relation the @ mark symbol give the relation, then we have to give the attributes. Here in this case now we use the composition; so, what are the attributes here?

Student: (Refer Time: 02:48) amino acid.

Composition, 20 different amino acids amino acid residues. So, we have 20 different amino acid residues and we have to say what the type of the amino acid again this is a numeric and this is the different composition then we what we have to do? So, that we have to explain; so in the curly bracket you have to give.

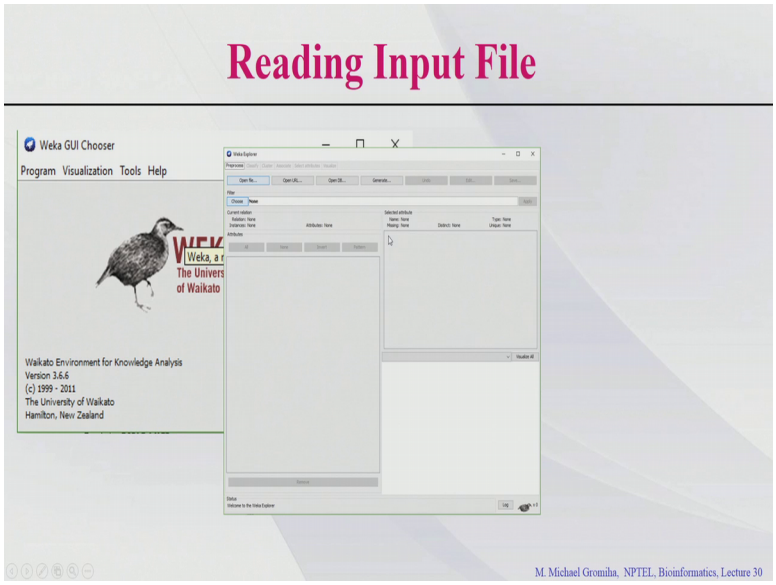
So, you have to classify here the main aspect is TMH and TMS; This is alpha, This is beta. So, they put class we need classify into alpha and beta then this is your properties or features what you have to provide. Then we have to give the data; what is the values for your dataset; so, we start with the @ data.

So, here then in next it starts with the data. So, what data you have to provide first we have 20 attributes; so 20 attributes or 20 compositions this is the amino acid

compositions for the 20 residues the same order. So, with the order we give here; so we give the same order in all the proteins and here this is the output because we know that this is alpha. We know that we put the values alpha, if you do not know you can put a question mark then that will assign. So, first we try to use a set of data what we have and see how they perform within the same dataset and how to group into training and test and so, on.

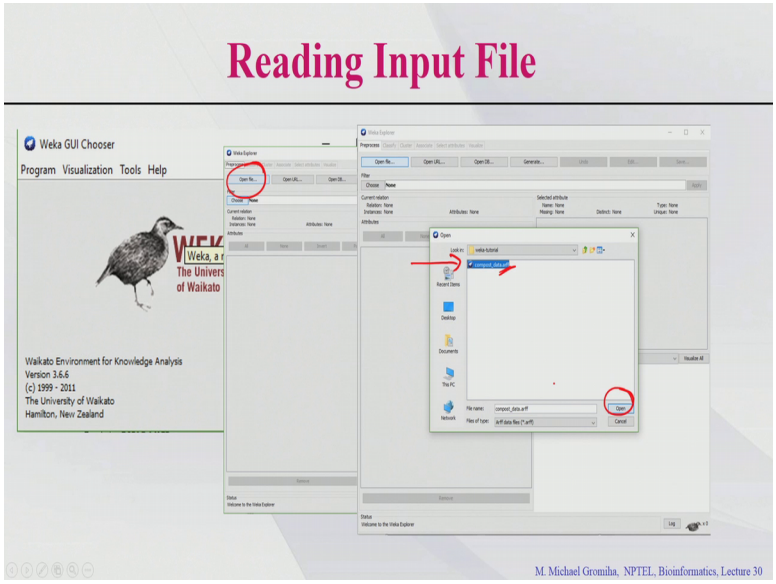
So, in this case you take all the composition of all the TMH protein and TMS proteins. And here you prepare the input file with all the data; so all data, now dataset is done. So, this is the arff format you have this is the attributes; this is the class and this is the instance data points, so now what you have to do?

(Refer Slide Time: 04:33)



So, if you click explorer then this will open a file; this window then here various options. So, first is we need to preprocess the data.

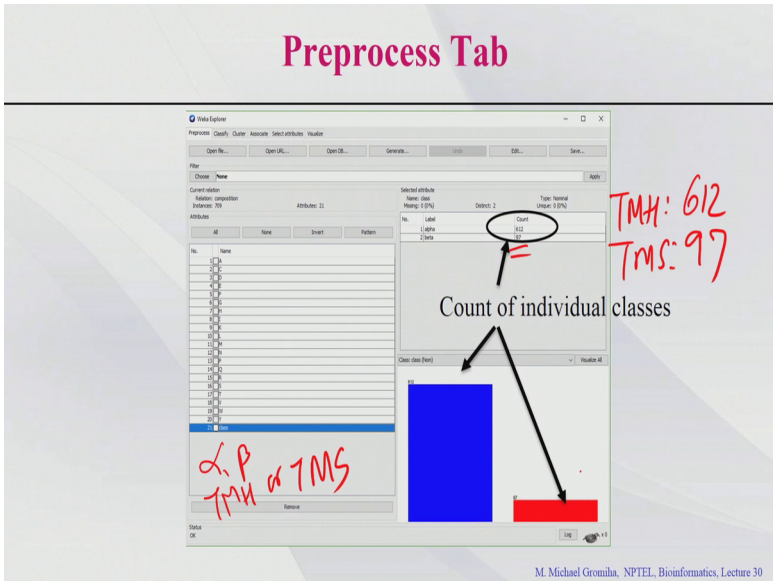
(Refer Slide Time: 04:49)



That means we need to input the data. So, we have to open file; so if you click on the open file this will ask where we need to get the data.

So, this is your computer then this is the location of the file; this in arff format. So, in if you click on open, we can open the file if you click this one and we will get this data here.

(Refer Slide Time: 05:06)

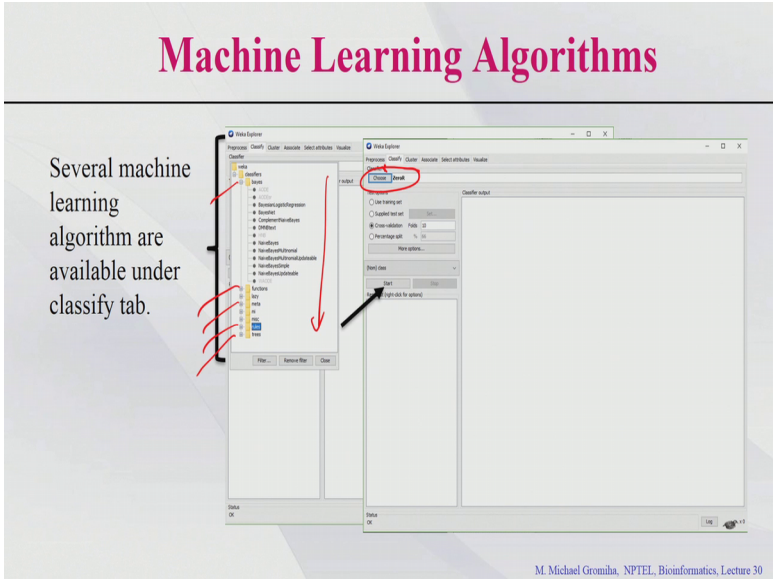


If you see the left side that you have in the under the name; we can see the 20 alphabets, these are 20 amino acid residues. These are the 20 input values given for each protein then 21; we gave the class, what is your class?

Student: Category.

This is alpha or beta or TMH or TMS. So, now we can see the count if you have the dataset we have 3000 alpha plus about 300; 400 beta; when we get the redundancy remove the redundancy, we get the non redundancy sequences. Now how many helical and transmembrane strands? 612 helices and beta is 97; so, TMS is 612 and TMS this is 97. This is the data size or these are instances you can visualize the instances; is equal to helix and 97 TMS you can read the data.

(Refer Slide Time: 06:17)



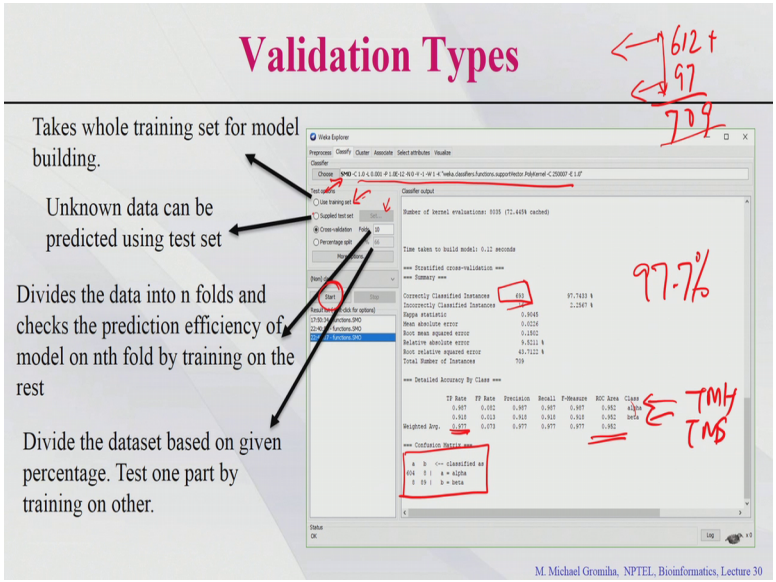
Now the next question is right, you have to choose the algorithm. The features we gave; the problem we set; we calculated the features, we gave the features; now we need to choose a method.

So, various methods available in the weka platform; so, broadly the classify as Bayes, or the functions or the meta and the rules, trees and so on. Under the bayes there are NaiveBayes, BayesNet and so on; if you go with the functions there will get the different logistic regression, other aspects. Likewise, you can see the neural networks, multilayer

preceptron. So you can get various, support vector machines, under various options. So, we will have more than 50 programs that are incorporated in this weka platform.

So, you can try various options and you can see the performance. Now we have the input data which was the input data and we have the functions which was the functions; that is also fine. So, now you have to start; so here the default option is zero because this is the default option given in weka platform. But if you want to choose; you have to click on choose then if you click on this choose it will list all the techniques and you can expand see any of this major classes. And you can select the which option do you want if you want to support vector machines, you can use the support vector machines.

(Refer Slide Time: 07:52)



If you keep the starts, but this will give you this result; for example, if you take this support vector machines this will give this is the SMO. So, these are the default values they used in a support vector machines and different parameters; I will show you now, different default parameters. So, now here there is various options available in this weka platform here; if we use training set this is the used training set what is the meaning of used training set?

Student: Self assessment.

If you give how many data we gave?

Student: 600 plus.

We have 612 plus 97; this will be 709 so you take all the 709 data for training and it will give you the data; they will give you the final result. Then if you have the supplied test sets; that means, you can train the weka for with any learning technique; for example, support vector machines with respect to this 709 data and if you have new data set for example, another 50 data. So, you can give it as test set supplied test set; if you click on supplied test set this will be open and then this will be enabled.

So, then I can click on the set; this will ask for the new dataset and from your computer terminal, you can give the new file or this will use that file as your test set. Then the third option is the cross validation: 10-fold cross validation. For example, it is 10-fold cross validation.

Student: Data (Refer Time: 09:35).

You can use the out of 709; they classify into 10 groups, then they use a 90 percent as training and the 10 as set. So, you do it for 10 times finally, give you the values; so take the 10 percent out here 10 percent out here for the testing and do it for 10 times. So, now, if you here if you give the 10; so they give the 10-fold cross validation.

Then also there is another option called percentage split; here you can see how many how much percentage you want to use for training and for the test. So, the default is 66 this why if you see many papers they uses 66 percent split because they do not change this just use the default values, now this you can change here, if you want to have the say 80 percent or 70 percent. So, here they mention 66 because two-third; this is why they put 66.

So, you can change this percentage split and you can try to evaluate in the performance of the method fine. So, when you do the cross validation and start click; so we can this is the start button; if you click it will give you the data. Out of this 709; so 693 are correctly predicted. So, the percentage accuracy is 97.7 percent and the error is 16 are wrongly predicted.

So, the error it is 2.25 percentage; now this is method is a good it is fine. Then can we see only the accuracy? No, but we have to check the sensitivity and specificity because it is the unbiased; this is the not the same dataset, you can see the different types of levels of data the 612 for the helix and the 97 for the strand that is not equally distributed.

So, in this case you can see the true positive; then one you this for the alpha, this is for the TMH and this is for the TMS. So, this will give you sensitivity and specificity for the case of the TMH this is 98.7 and the 91.8. So, if you take the weighted average this is 97 percentage that is fine. So, always you have the other information regarding the false positive rate precision recall f measure ROC. So, ROC is 0.95; so how the graph looks like?

Student: (Refer Time: 11:55).

Very high; look like the straight line that mean this is very high 0.95. So, in this case your given area under the curve that is very high. So, this method is reliable you can do that, so you can see the confusion matrix you this is the confusion matrix what is the meaning of confusion matrix?

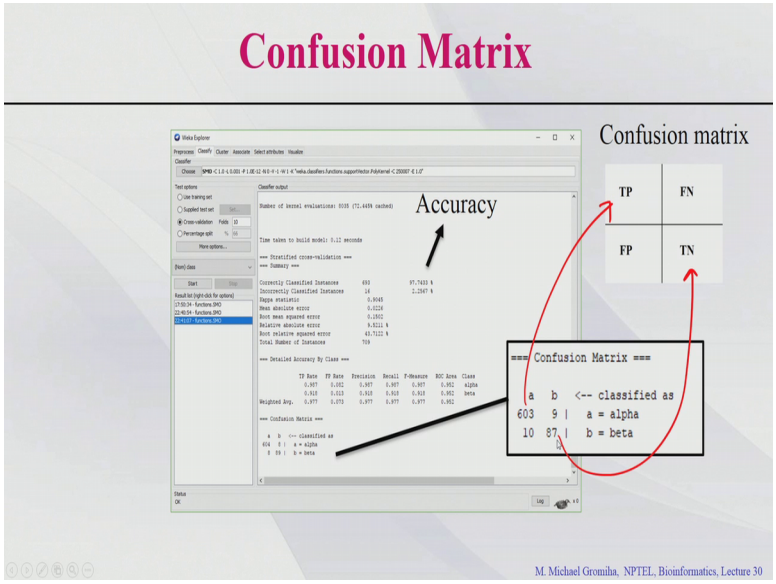
Student: (Refer Time: 12:13).

It can give you true positives, true negatives, false positives and false negatives; if TMH is positive what is the meaning of true positives? TMH predicted as?

Student: TMH.

TMH; this is the true positive you can see confusion matrix.

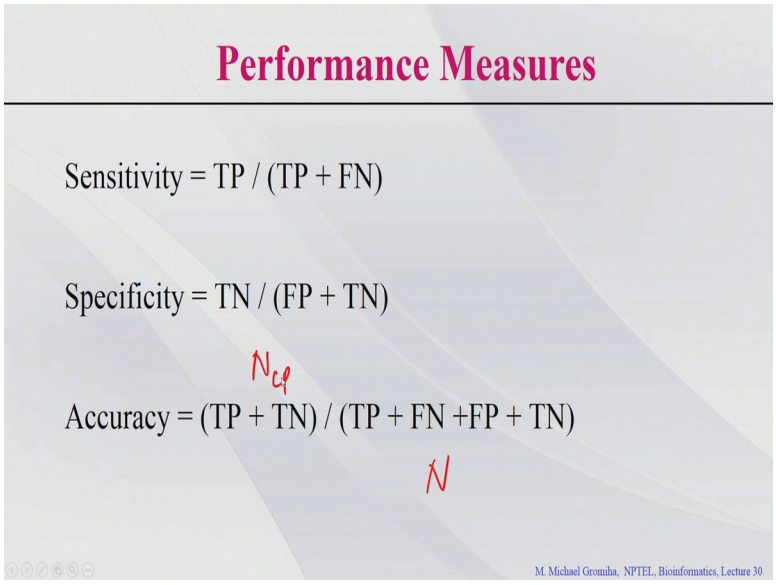
(Refer Slide Time: 12:26)



So, this is true positive you can see this is the true negative. So, this is the false negative and this is the false positive. So, if you see this one then you can get the sensitivity true positive 603 divided by total positives this is the 702. So, in this case this is the sensitivity; this is here in this case only 87; so, this is specificity is less; this I also I discussed one of the issues with the machine learning.

So, you can see this is highly trained with this positive dataset, but in this case it is also good that the negative dataset is also properly trained because the specificity also 91.8 now this case it is fine good.

(Refer Slide Time: 13:15)



So, this is the formula to calculate the sensitivity specificity and accuracy; you can calculate sensitivity by true positive by the true positive plus false negative. Specificity is true negative by false positive plus true negative, the accuracy is total positives with the all number of data this is all number of data this is the correctly predicted data.

It is number of correctly predicted data you can see that. So, now, if you want to see here what are features we used here?

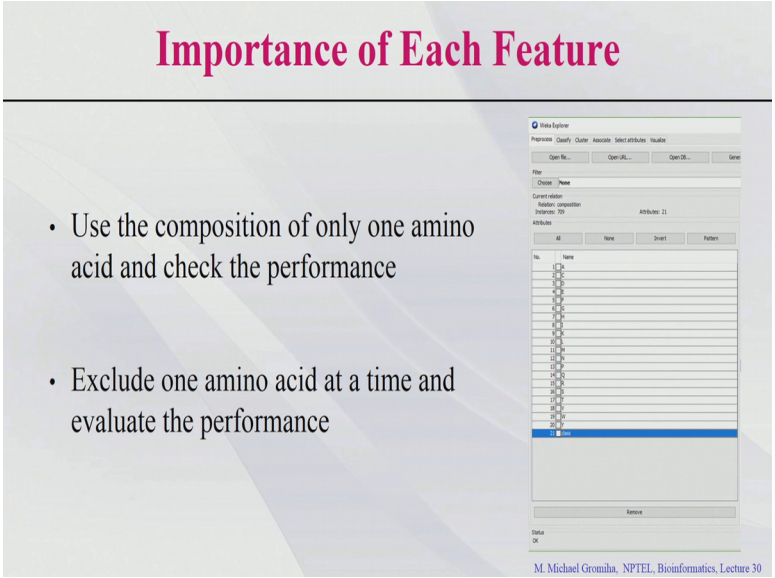
Student: Amino acid.

Amino acid composition; how many composition we used?

Student: 20.

20 composition we have 20 amino acid residues your 20 compositions; now the question is whether we need all the 20 amino acid composition to get this values? Because as I discussed earlier, we should minimize the features for any classifications even the QSAR we also discussed. We should minimize this features in case you can reliability will be high; the error will be high, the error will be less. So, in this case we see whether the composition of all the residues are necessary or we can eliminate some of the amino acids. If amino acid reduced; if you do not make any change then we can eliminate this amino acids; you can reduce this number of features.

(Refer Slide Time: 14:25)



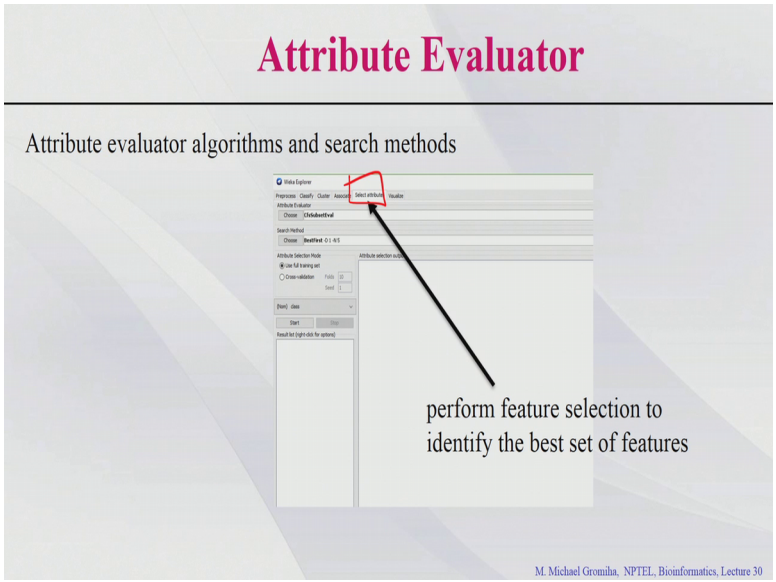
So, in this case you eliminate one amino acids and see whether using other 19 amino acids the performance is similar or any change. If there is any change or similar one then you can discard; if you drastically reduce the performance then that feature is very important; you have to keep. Likewise, you can do all the 20 you can do one by one (Refer Time: 14:48) one amino acid and then see which amino acids are not very important for classification.

Then you will have 19; for example, if you remove 1 or 2 then we have less then again you try the other one. So, again among 19; you can try to eliminate one; so while reducing this features, you can optimize the features manually you can do that. But weka there is also another option; it will also automatically select the features, but in anyway if

you have to less number of features; you can also try manually doing all these things those who can optimize the features.

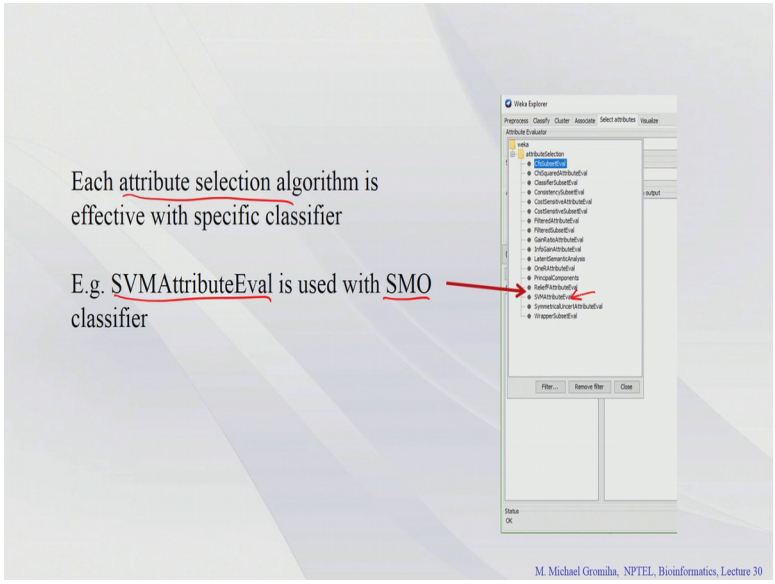
There is one way second one we can use only one amino acid say just alanine; what is the performance of alanine where is 50; 50 percent or it is 80; 20 percent or 70; 30 percent and doing all these 20 features, we know the important features which one which one have the highest performance; when you combine this features and then also you can enhance the performance say the highest performance; with reduce number of features this is helpful for the reducing the number of features.

(Refer Slide Time: 15:47)



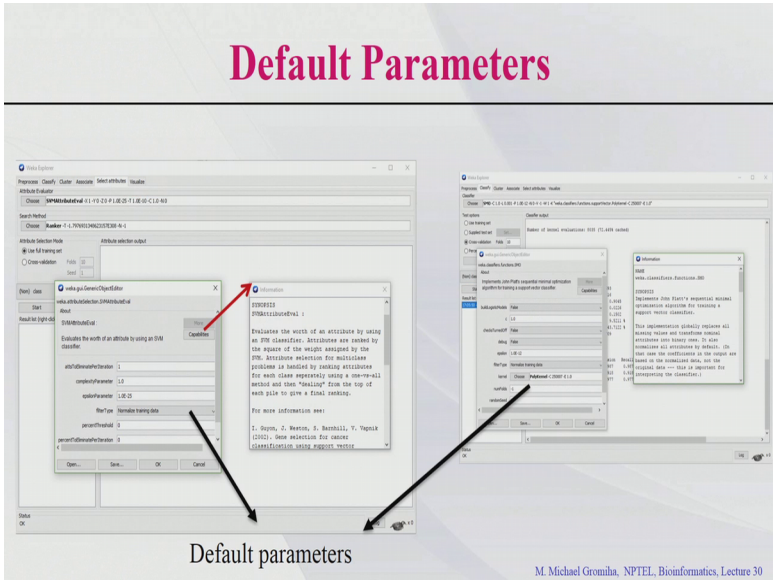
So, in weka also we have the attribute evaluator we can manually do that or you can use the weka you can see the attribute select. This perform the feature selection to identify the best set of features, if you do this; this will also give you one option these are the probable features which can give the best performance.

(Refer Slide Time: 16:06)



So, for in this case each attribute selection algorithm is effective with specific classifier. So, for each classifier they used different algorithms to optimize these features. For example, if you take a support vector machines they use the SVM attribute evaluator; this is here to optimize these different features.

(Refer Slide Time: 16:29)

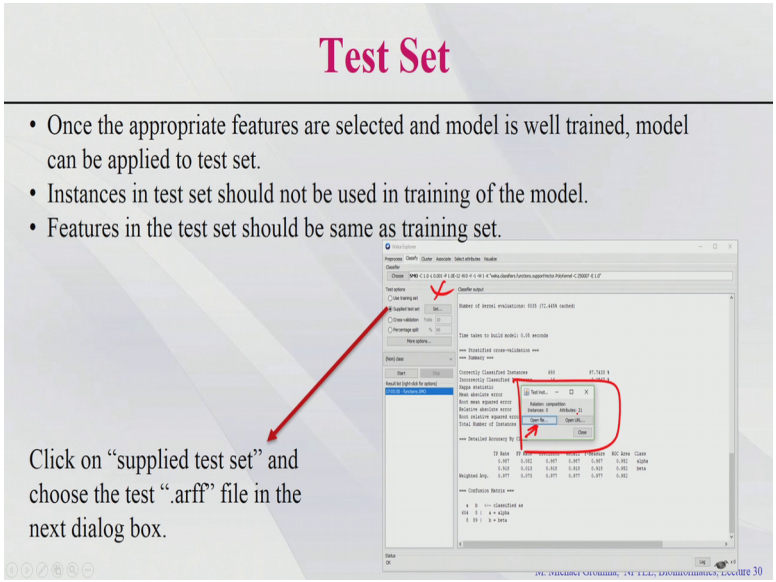


If you do this; so there are various options you can give that. So, here you can see this is option this is the SVM attribute evaluator. So, what are the various capabilities? For example, if complexity parameters or the different types of filtering; how to normalize

the training data and so on. Then for the classifier also we have various options; we can use this different file types or what are the different kernels we want; here we use the polykernel type also you can see all the details in the weka platform.

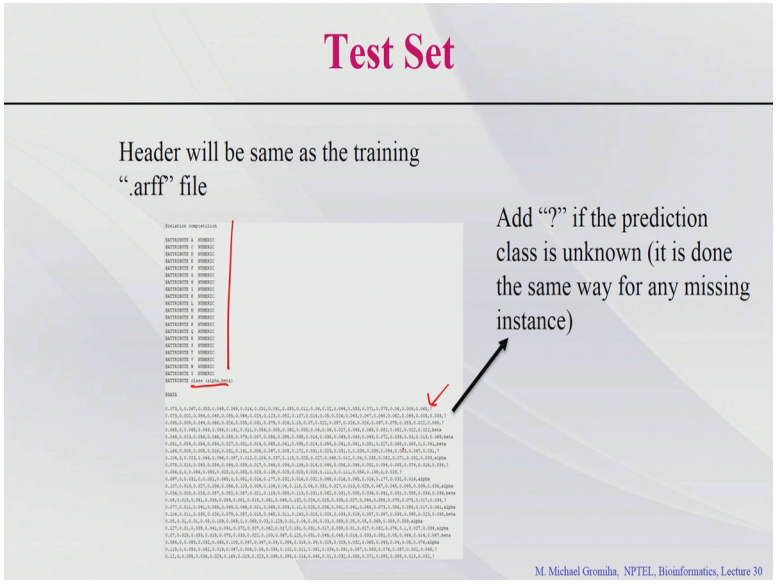
So, for any machine learning algorithm use; you have to first check what are the various features available. For example, if you use a neural networks you can assign your hidden layers; so all these information they set to default value trained with some normal data, but with respect your dataset. So, fine tuning this default parameters we can improve the performance; not just relying the default parameters given in weka. If you fine tune the parameters, you can improve the performance at least 3 to 5 percent.

(Refer Slide Time: 17:41)



So, now we can see the test set that I discussed earlier how to get the test set. So, here this is the supplied test set if you give click on this set, then this box will open. So, in this case open file here; so we can give the file and then the supplied test set, you can get the data.

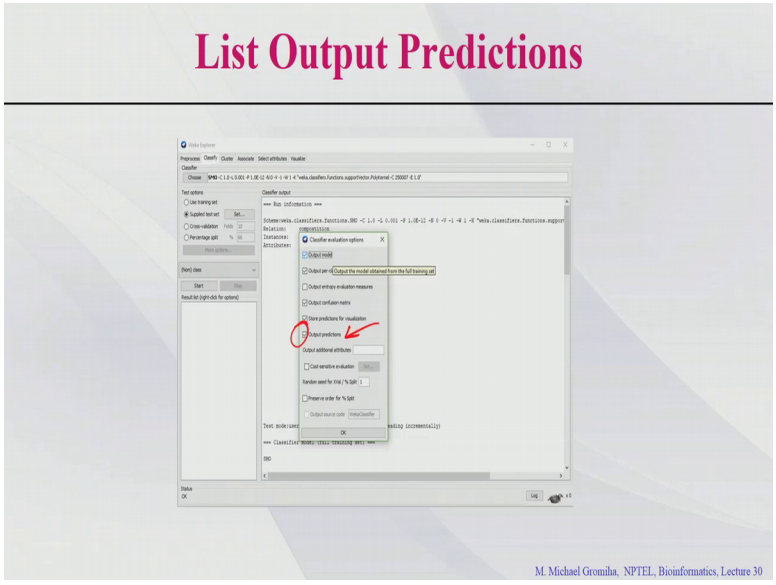
(Refer Slide Time: 18:04)



So, now here this is the data set for the test set; in this case we use the same composition the features these are the composition we use 20 amino acid residues.

The same class data if you see the here; if you do not know the test set you can the question mark. Now question mark means the prediction is unknown; so in this case the weka will be predict this helix or strand. If it is known alpha and beta it can also evaluate this is the positive or negative, where it is correctly predicted or it is wrongly predicted.

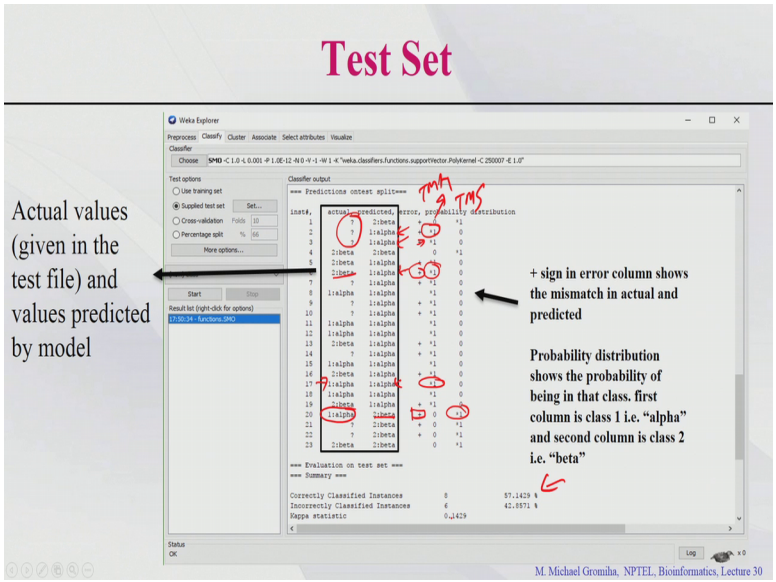
(Refer Slide Time: 18:38)



This is the output; so it is also possible to list the output predictions for example, here there are various evaluation options. You can give the output model or you can we can output the confusion matrix and also you can put the output predictions. If you click on this output predictions, for each case this will give you the data; what is the probability of the particular protein to be in TMH or in TMS?

For test set data it will give you the probability and based of probability it will decide. If you know the data, it will tell the probability for TMS and TMH and then see this is the positive or it is error; I will show the data here this is the case.

(Refer Slide Time: 19:19)



So, if this is the we do not know it is a question mark. So, this is the predicted as 2 this is the beta for example, if you see the probability 0 for TMH and probability is one for the TMS; so, this is beta.

Second case, if the probability is one for the TMH; so it is predicted is alpha. Third one is also one this is for alpha; sometimes we get the different numbers not 1 and 0; 0.6, 0.7; in the decimal also will get based on the higher values, it will classify either alpha or beta. So, if this is the this is already known; this is the alpha in this case this is predicted as alpha then it is fine; if it is a known is beta this is one this is predicted as alpha this is wrong.

If it is wrong it mention as plus here; in this case if you what was see the plus signs here? Why are we have the plus signs? This all wrongly predicted; for example, in this case this is a alpha is actual case, but the predicted case one is in the second case beta. So, it is predicted as beta; so it is a wrong; so put the plus sign, in this case this is alpha with this number this is one is here. So, predicted as alpha; so that is fine.

Finally you can get the final data this case the classification performance is 57 percentage; this is good? This is not good. So, when in this case we have to go back in the training set we need to change the features; for example, you can include the hydrophobicity as one of the features you can from this knowledge you can add more features and use a training, use a cross validation and use a test if you are able to predict then it is fine; otherwise you need to work on this features and you need to manipulate the giving the feature selections. And finally, as long as your convinced you need to train this machine learning to get the highest performance.

So, in summarizing; so what did we discussed in this class?

Student: Application.

Applications for example, classifying transmembrane helical and transmembrane beta barrel proteins. Mainly with this statistical methods as well as the machine learning techniques, in statistical methods we can compare the composition of the standard set. For this we need to develop your data set and calculate the composition and compare with the unknown ones and based on the deviation; deviation obtained from the hamming distance or the euclidean distance and you can decide. Then you may use some error function to decide what is the error function which gives a best performance.

Then go with the machine learning; what are different types of machine learning we discussed? Supervised.

Student: (Refer Time: 21:59).

Unsupervised learning; what is a example for supervised learning?

Student: Neural networks.

Neural networks, support vector machines for the unsupervised learning?

Student: (Refer Time: 22:05).

k means clustering; so, different types of machine learning techniques then we discussed about the weka; what is weka?

Student: (Refer Time: 22:13).

It is a platform.

Student: Machine learning.

Which contains with different types of machine learning techniques, we can use machine learning techniques as well as the feature selection algorithm, you can merge in the weka platform. So, for example, if you want to use weka; so, we want to prepare the input files; what are the accepted input files? Which formats? arff.

Student: (Refer Time: 22:35).

Plus csv; you can use it. So, how to prepare the arff file?

Student: Attribute.

First give the attributes, then give the values and then the class; then finally, give the data then this will give. So, what are the various options available in weka for the evaluation?

Student: Sir if I use.

You consider training set or the cross validation or the.

Student: (Refer Time: 22:56).

Split sampling also you can use the external set; when you run the weka, it will give you the confusion matrix; what is the meaning of the confusion matrix?

Student: True positive.

True positives; true negatives, false positives and false negatives using this confusion matrix you can calculate sensitivity specificity and accuracy; also weka directly you can get the values. Along with precision as well as the ROC error; so now we can use this information to use to predict or to discriminate or any different classes of these proteins.

Also here we have the attribute selection algorithms; this will also select the preferred attributes to give the best performance.

So, there are various applications in weka; if you see the lot of papers they published using weka software. Also another important thing is instead of using the default parameters set in weka; for any programs support vector machines or neural networks and we can change the default parameters and see.

So, then how they change the performance; you can improve the performance at least 3 to 5 percent. So, there are plenty of applications some of them we discussed in the last class all these applications, we can use; the weka platform and select the best machine learning techniques to achieve the highest performance. So, in the next class I will summarize all the different techniques or the algorithms or the concepts; we learned in the past several classes, I will summarize the all the information in the next class.

Thanks for your kind attention.