The workshops conducted allowed for the use of different types of classifiers and subsequently testing their performance.

We started with an initial data set containing data on 2951 subjects: these were divided into construction sets and test sets.

The construction set was further divided into training and validation set using three different methods: in the first method, the random method, one starts with the non-discretized construction set and assigns to the training set 60 per cent of patients from the least represented class and an equal number from the other class; the remaining patients will make up the matrix of the validation set.

In the second method, based on hierarchical clustering, one starts with the construction set normalised by the minimax scaling method and then divided into two submatrices, one for each class.

For each of these submatrices, a dendrogram is generated and after finding the best cut, each submatrix is divided into clusters. Finally, for class zero, 60 per cent of patients in each cluster are extracted randomly and allocated to the training set; the remaining 40 per cent are placed in the validation set.

for class one, on the other hand, the number of patients proportional to those extracted for class zero is extracted for the training set, with the remainder in the validation set

In the third method, that of self-organising maps (so-called SOMs), one starts with the normalised construction set, which is divided into two sub-matrices, one for each class. Then, after estimating the parameter n1, three som of different sizes are constructed for each class, obtaining a total of six networks

these networks are then trained and a map of neurons is produced for each one containing the number of wins for each of them. in order to identify the best cut in the associated dendrogram, these maps are analysed with a view to identifying groups of contiguous neurons. then after identifying the winning neuron for each element, the clusters are obtained according to the winning neuron element association. the centroids of each cluster and the inter-cluster distance are then calculated.

Finally, the best som size was identified for each class as a function of inter-cluster distance.

Again,a training set was created by extracting for class zero 60 per cent of patients from each cluster and for class one a number of patients proportional to those extracted for class zero and the validation set containing the remaining patients.

In lab 2, a genetic algorithm was used to select the most relevant variables in the case of codified variables from the discretized construction set. then the non-discretized construction set was divided into training set and validation set. from these matrices, normalised by the minimax scaling method, a genetic algorithm was applied to select the most relevant variables in the case of codified yes no variables. finally, a KnN classifier was constructed and applied to classify the patients of training validation set in the case of both codify variables and yes no variables.

in lab three, the non-discretized construction set was again divided into training set and validation set but using two different methods: hierarchical clustering and the SOM method. in both cases a KNn classifier was then constructed to classify the patients of the training and validation set. finally, the confusion Matrix obtained with the three methods just described were compared after the KNn classification.

lab four starting from the subset of features, obtained by encoding yes novales, two classifiers were constructed based on the perseptron multiplayer by means of the creation, initialisation and training of supervised neural networks. one classifier was based on the manual choice of the best structure out of the 8 proposed, the other on the selection of the best structure by means of ga, which provided a more relevant subset of features and the binary encoding of the structure.

In workshop 5, finally, starting from the set of features obtained by encoding variables, the membership fuction was constructed for each selected feature: using the latter, we then proceeded to the construction of the rules that made it possible to define the classifier based on fazzy logic.

2)

this slide shows the Confusion Matrix training set and validation set for all the classification methods used. there are three classification methods: the knn which classifies an element on the basis of the similarity of elements in its neighbourhood

ANn based on the multiplayer perceptron is a supervised learning algorithm and fis so fuzzy logic, through class

In general, it can be seen that the knn and ANN classifiers are able to assign each element of the two sets to a class, albeit with some errors.

the fis classifier, on the other hand, reports unclassifieds.

In this slide, to assess the goodness of the classifier, we used the % of correct classified on the class. This parameter allows us to decide whether the classifier has classified the patients correctly with respect to their class.

In general, it can be seen that the classifiers do not perform fully satisfactorily: we are presented with an unbalanced % of correct classified knn towards class 1 patients; this situation represents a potential risk if the classifier is used without knowledge: we risk missing out on sick patients, classified as healthy, and jeopardising their health status even more.

As far as ANNs are concerned, the situation is rosier: one has been able to construct better classifiers that could be suitable for use in less risk-prone situations, e.g. screening tests.

A separate issue for the fis classifier: it returns too many n.c., rendering it unusable.

To remedy this, dendrograms were cut at a lower altitude, with the aim of reducing intra-cluster variability and inter-cluster distance, but despite this and the construction of as many as 76 rules, the results obtained are still unsatisfactory.

Another attempt that could have been made would have been to widen the extremes of the membership functions by about 10%, a solution presented by the professor in the lecture based on her experience with such classifiers.

the different classifiers were used on the test set matrix containing never-before-used patient data to test their reliability and generalisation capabilities.

This slide shows the confusion Matrix related to the testing process of each classifier

We can see, comparing the classified correct % of validation and test sets, that the algorithms show a good generalisation capacity, determined by the insignificant decrease in them. This good generalisation indicates a good tendency to recognise and classify even elements not used in their construction.

THANKS FOR YOUR ATTENTION