**REPORT ANALYSIS CAR EVALUATION DATASET**

The dataset considered is a dataset on evaluation of the cars. The instances in particular are 1728.

The model evaluates cars according to the following concept structure:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PRICE | MAINTENANCE COST | NUMBER OF DOORS | CAPACITY | LUGGAGE BOOT | SAFETY | DECISION |

Number of Attributes: 6

Attribute Values:

buying v-high, high, med, low  
maint v-high, high, med, low  
doors 2, 3, 4, 5-more  
persons 2, 4, more  
lug\_boot small, med, big  
safety low, med, high

Missing Attribute Values: none

The decision can be: ‘unacc’(unacceptable), ‘acc’(acceptable), ‘good’(good), ‘vgood’(very good).

The dataset is structured with one instance for each possible combination of conditional attributes, then the possible combination of all conditional attributes is 1728. All the analysis from this point is done with the application developed in pyhton code.

If all the attributes are considered obviously the roughness is 100%, the roughness is the accuracy of the data and is obtained with number of elements in the lower approximation over the number of elements in the upper approximation. If all the attributes are considered and the concept is fixed in decision= ‘unacc’ in the positive region are contained the 70,023% of the data. If the same thing is done with decision= ‘acc’ the positive region contains the 22,22% of the data. For the decision ‘good’ is 3,99% and for decision ‘vgood’ is 3,76%. Then we can say that the dataset is distributed:

* 70,023% of cars has an unacceptable evaluation.
* 22,22% of car has an acceptable evaluation.
* 3,99% of cars has a good evaluation.
* 3,76% of cars has a very good evaluation.

For a good analysis is good if is focused on the decision unacceptable or acceptable, for other two evaluation the percentual of data is very less.

After an accurate analysis the alpha and beta are fixed in 0.7 and 0.3 but is considered also the values of 0.6 and 0.4 .

The following analysis is done with the sequential three way method(based on Probabilistic Rough Set Theory) , the attributes considered more frequently are safety, maintenance and price.

***ACCEPTABLE EVALUATION***

Start with an analysis on evaluation ‘acceptable’, and start with only attribute safety that can assume three different values(low, med and high), than the IND[Safety] possible are three.

In the positive region are present 0 elements, while in the boundary region we found the 66,67% of the data and then in the negative region are present the 33,33% of the data. With only one attributes is almost impossible come to some conclusion but is possible say that also if a car has a safety high and surely is combined with other attributes that can be positive the evaluation is not acceptable surely .Remember that alpha in this case is 0.7. In this case the roughness is 0%.

Go to do other analysis adding the conditional attribute Maintenance (that can be low , med , high , vhigh), this is referred to the cost of maintenance , only to be clear. The IND[Safety, Maintenance] classes possible are 12. Adding this conditional attribute the data contained in the negative region are greater than the previous analysis and then reciprocal boundary region contains less data, the change is important, the 8% of the data. The positive region is empty yet. But what happens if we change the values of alpha(0.60) and beta(0.4). Now there is a significative change in the region. The 33%of the data contained in the boundary region is now contained in the negative region. Now in the boundary region is present the 8% the data and in negative the 91%. This means that with this conditional attributes only 8% of the data is contained in a class that has a probability between the 40% and 60% to have elements that belong to the concept.

Go deeply in the analysis and add the conditional attribute ‘ price’ with an alpha(0.70) and beta(0.3). The negative region is populated by the 62.5% of the data and in the boundary we found 37.5% of the data. The positive is empty yet. That means if are considered these three attributes 0% of the data contained in the respective class has a probability to have elements in the concept that is equal or greater than 70%. Try another time to change alpha(0.6) and beta(0.4). The data change significatively . Now the boundary region is empty, that means if are considered safety, maintenance and price no classes have a probability between 40% and 60% to have elements in the concept. The 37% is distributed in positive and negative region. In particular in the positive region now the data contained is the 20.83% of the data and negative had an increase of 16.67%. The 20,83% of the data contained in the respective classes has a probability to belong to the concept that is between the 60% and 70%. In this case the roughness is the 100%, that means upper and lower are equivalent.

For discover other things try to add an other conditional attribute that is the number of doors(2,3,4, 5more). Now the possible IND[Safety, Maintenance, Price, Number of doors] classes are 192. With alpha(0.7) and beta(0.3) the negative region contains 61,45% of the data and boundary the 38,54% of the data. Any of the classes has a probability greater or equal than 70%. If we try to change alpha(0.6) and beta(0.4) the boundary region loss of the 25% of the data that are distributed with 15,625% of the data in the positive and about 10% in the negative. The 15,625% of the data with these attributes has a probability with the respective classes to belong to the concept between 60% and 70%. We can say that with number of doors the positive region contains less data, than the number of doors has a negative impact on an acceptable evaluation of the car combined with other attributes. The roughness is now 53,57% , about the 50%of the data contained in the upper is also contained in the lower. If try to add other attributes the result is not a good analysis, the ind classes are high if compared with the number of observations.

***UNACCEPTABLE EVALUATION***

Now the concept is fixed in evaluation unacceptable. Start again considering only the conditional attribute safety(low, med, high). The IND[Safety] classes available then are three. With an alpha(0.7) and a Beta(0.3) the 33,3 % of the data are contained in the positive region while the 66,7% is in the boundary. The data is equally distributed and we can deduce that are because for each level of safety there is a different probability. It is possible say that almost surely when the safety is low the decision of the car evaluation is unacceptable combinate also with the other attributes, and in particular the 33,33% of classes have a probability that contains the element in the concept greater than 70% . But what happens if the alpha(0.6) and beta(0.4) change. In this case one classes passes from the boundary to the positive that now contains the 66,67% of the data. Is possible say that if is considered only the safety 2/3 of the data has a probability to have an unacceptable evaluation that is greater or equal than the 60%.

Try to add the conditional attribute ‘price’ and then now the classes are IND[safety,price]. The number of possible classes are now 12. The data is equally distributed among positive and boundary, in both there are the 50% of the data. The roughness is now 50%, and then the middle of the data in the upper are also in the lower. Is possible speculate that the data in positive region are the combination of safety low-med and price med-high but is only an hypothesis. If alpha(0.6) and beta(0.4) are changed the data change. Now the boundary contains the 25% of the data, it loss the 25% of the data that are now distributed in other two region and in particular the 16.7% go in the negative region and the 8.3% adding the positive region. The 16.7% of the data with the conditional attributes of safety and price has a probability between the 30% and 40% to have an unacceptable evaluation. The 8.3% of the cars if are considered only safety and price has a probability between 60% and 70% to have an unacceptable evaluation.

Add the conditional attribute ‘maintenance’ ,that is fundamental for the analysis . Le IND[safety, price, maintenance] now are 48. Con alpha(0.7) and beta(0.3) the positive region contains the 45,83% of the data while in the boundary find the 54.17%. Roughness is now equal to 45.83%. If is compared with the previous analysis the 4.17% of the data passed from the positive to the boundary. Maintenance cost is a fundamental attributes that affects on the unacceptable evaluation, with Maintenance there is a bit uncertainty more. But also in this case try to change alpha(0.6) and beta(0.4). The variation is significative . The boundary region is now empty , 54,17% of the data are contained into is now distributed in positive and negative and then the 54,17% of the data has a probability with his class to belong to the concept that can between the 30% and 40% or between the 60% and 70%.

The 37.5% of this data is now contained in the negative region. The 37.5% of the cars considering the attribute safety, price and maintenance has a probability between the 30% and 40% to evaluate unacceptable. The 16.67% has a probability to been evaluate with an unacceptable evaluation that is between the 60% and 70%. The roughness is now 100%.

***FINAL CONSIDERATIONS***

Obviously these probability must be considered in the complex of the belonging to a specific indiscernibility class).

Adding the capacity of the auto could be misleading. After a specific analysis is deduced that when the capacity of the car is equal to two the evaluation of the car is any case unacceptable , with all possible combinations of conditional attributes. This must be analysed deeply, but for this dataset the car with the capacity of two person are overvalued.

As already said the dataset contains a single possible combination of all attributes and the evaluations are strongly unbalanced towards the unacceptable evaluation.

On this surely affect the capacity of the car when two that conditions the 33,34% of the data, the 1/3.

Attributes more important are safety, maintenance and price .

If is considered only the safety attribute with an alpha(0.7) and beta(0.3) no car has an acceptable evaluation. Must be considered the three fundamental attributes and change alpha and beta to obtain an acceptable evaluation , the 20.83% that has a probability between the 60% and 70%.

Maintenance cost and price are two of the most important attributes, when these increase the decision go versus the unacceptable evaluation, depends the values of the other attributes.

***CODE ANALYSIS***

The code is divided into two parts , the main and the 3WD package that contains all the functions to execute the sequential three way decision with the probabilistic rough set.

In the main the code is specific for the dataset analysed, here are read the csv file that contains the dataset and the file that contains the order of the columns for the sequential. The only two parameters required are the values of alpha and beta. The output is given by the calling of the only function of the main, the Sequential function that permit to analyse the dataset in the order that the user prefers. The sequential function is structured more simply, is created a list that contains the names of the columns that must be to considered. The first is always the decision column. In the cycle are called the functions of the package and are printed the columns considered, how many different indiscernibility classes are present for each iteration, the percentual of data in each region and the percentual of the roughness. This is all for the main analysis. Jump to the package.

In the package there are six different functions. The more simply is ‘Concept’, that has two parameters, the decision columns and the attribute that the user want consider, after are returned only the indexes of the rows that correspond to the concept.

In the indiscernibility function is considered the object that is passed into the parameters and are created two empty list. Is considered after the list of columns and is deleted the first columns because is the decision columns, this is way in the sequential function of the main the decision column is appended as a first. Is important know this. Then all the rows that have the same attribute for the columns that the application is considering in the specific moment are grouped in the same list and after only the indexes of this are saved into. This function return all the list with all the indiscernibility classes.

Positive, Negative and Boundary functions are more similar. For each class of indiscernibility is calculated the probability computing the number of the elements that are present also in the concept over the number of elements of the class. Are positioned into the positive region if the probability is greater or equal than the alpha value, in the negative if the probability is less or equal than beta value or in the boundary region if the probability is among the alpha and beta values.

The last function is the roughness that calculate the accuracy. In this case are also calculated the probabilities, but for the lower and upper approximation. For the lower are considered the probabilities greater than alpha, for the upper greater than beta. Then the roughness is calculated the number of elements of the lower over the number of elements of the upper.