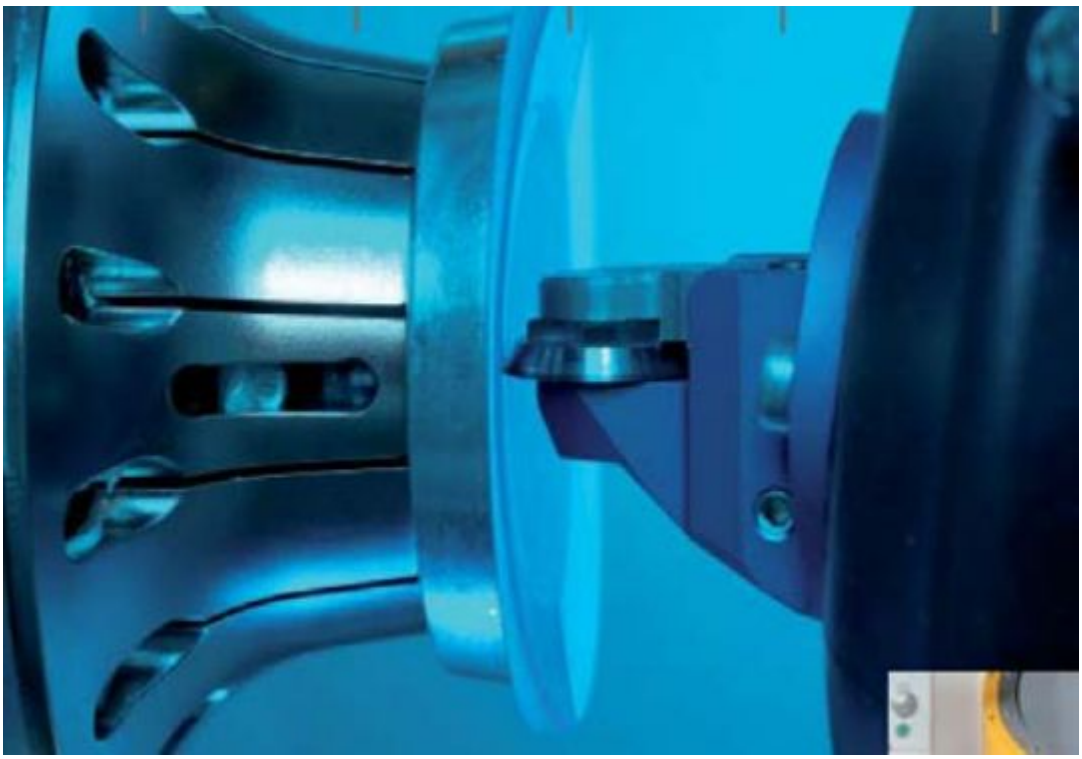
X MODELLING WEEK

QUALITY CONTROL OF OPTICAL LENSES



Students: Richard Burton
Alicia González
Elisa María Da Conceiçao
Carlos Perales

INDEX

1 .- INTRODUCTION

When many products can be made easily, a way to measure its quality should be set up in order to distinguish defects in the fabrication of the products. However, this task is heavy and often slow. How can we measure all the products fast enough to sell them? Here is where mathematics enter into the equation. Automatization of quality controls is needed to set up in order to make the task faster. However, we need to teach a machine how to distinguish between a good product and a bad product. There is when mathematical models, specially machine learning, is important.

Not a long time before, quality control was accomplished manually. Some technicians are employed to measure if a product was good or bad. However, this task was very slow, and not all the products were measured. A statistical approach was needed.

So, how can mathematics help into quality control? A mathematical model can be established in order to classify a product, by some of its characteristics, in good or bad. It works by statistical classifiers, also known simply as classifiers. These methods take information about a set of samples, called training set, with the features we want to distinguish from a product and the label. In our case, an optical lense is considered either “good” or “bad”.

So, a classifier tries to find the pattern between the features and the labels we imposed. That is because the training of the classifier, and the choice of the model in which the classifier is created, is so important. What is more, keeping the same training data and the same classifier, the the classification will be consistent in time. It will not change, regardless the circumstances.

In this report, the product to be treated is optical lenses. With a set of labeled data, we divided in training set and a test set, in order to calculate the accuracy. The division is remade several times in a processed called 4 cross fold validation. This technique of validation, basically, means that 4 exact division of the labeled data are created, and in each iteration one set works as test set and the others as training data for the classifier. The classifier is created by using the training data, and returns a prediction of the label a new sample should be considered.

Several classifiers have been created, but, before that, data have been explored in order to find the appropriate classifiers.

2 .- DATA EXPLORATION

The quality measure is established based on two experiments. From those experiments, the features to create the mathematical models are extracted. The way this is done in each experiment is by measuring properties of the experiments in a sensor, formed by 23x23 pixels.

Firstly, mean power is tested on an optical lens. The information the sensor collects is compared with the theoretical result of a perfect lense. From this experiments we get 23x23 = 529 features. The second experiment consist in a measure of the aberration the lense could get, from which other 529 features are extracted.

In the end, each sample has 1058 features, plus the target, “good” or “bad”. Process of the data in order to create classifiers can be hard. However, data can be compressed. We have tried two sets of data. One of them contain the data with all the features. The other just 4 features; whilst the first couple of them are the mean and the standard deviation of the first experiment, the second one is from the other experiment.

Essentially, the two sets contains the same information, but the second one is condensed (a summarization). It could have losen relevant information for the classification, that is why, if the computer performance allows it, both sets are tried.

In the start-up phase of the present study, an exploratory data analysis has been carried out for the four estimators (hereinafter referred as X) and the target in order to gain insight into the data. The aim of an identification of the data structure and potential relationships among variables is to provide significant support when deciding an appropriate algorithm to classify the lenses.

A display for each one of the four variables considered of the frequency histogram by each target value (good or bad) has shown the existence of certain common grounds between them.

- Lowest values are more likely to be associated to approved lenses and it is reversed as the values increase. This observation is in line with the dependent variables involved because they are referred to the mean or standard deviation of error values (differences between the theoretical values and the measured ones).
- The large overlapping of the histograms reflects the gradual shift in assessment towards to a more likely rejection as the values increase. In this respect, it is noteworthy the extreme situation in X_1, with a very large intersection region of both empirical distributions.
  According to this finding, none variable would be enough on its own to obtain a good predictive model for classifying.
- Data have high positive values that can be identified as outliers. All these values are associated to cases of rejected lens. This should be taken into consideration when deciding the classifier to prevent a spurious effect of these extreme values.
- Empirical distributions from the sample of rejected lenses (and, for the variable X_1, also from the sample of rejected lenses) are markedly skewed to the right. These also have a higher dispersion than the ones from the accepted lenses even when the extreme values had been removed in the analysis.

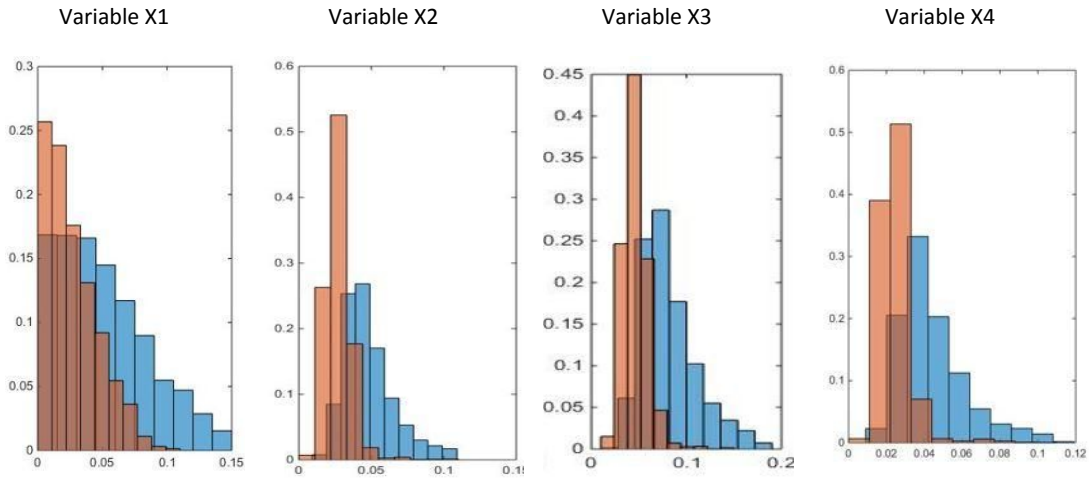


Figure 1 – Frequency histogram by target value for each one of the four variables (mean and standard deviation of the error made with respect to the theoretical values for the two dimensions measured) **(reduced sample**)- It has been removed the values above the percentile 95. Red histogram is referred to the accepted lens (good) and blue histogram to the rejected ones (bad).

Graphs by pairs of variables, displayed below, have allowed a fast overview of the relationships between variables and with the target. Looking at the overall sample, plots show that the above mentioned *extreme values* for either of the two variables considered could have certain influence in the correlation between them. It is also to be noted the distinct characteristics of the variable X_1 in relation to its extreme values.

Therefore, whilst the remaining variables identify the same cases with their most extreme values (i.e. allocated along or around the diagonal of the scatter plots), the values of these cases are not classified as outliers for the variable X_1, or at least they are appreciably less extreme. This comment is supported by the plots of X_1 versus the other variables (see Figure 1) that show the highest values for one of the two variables are close to the corresponding axis. According to the latter observation it was supposed that variable X_1 could provide a high gain of information to the model in case of working with the overall sample.

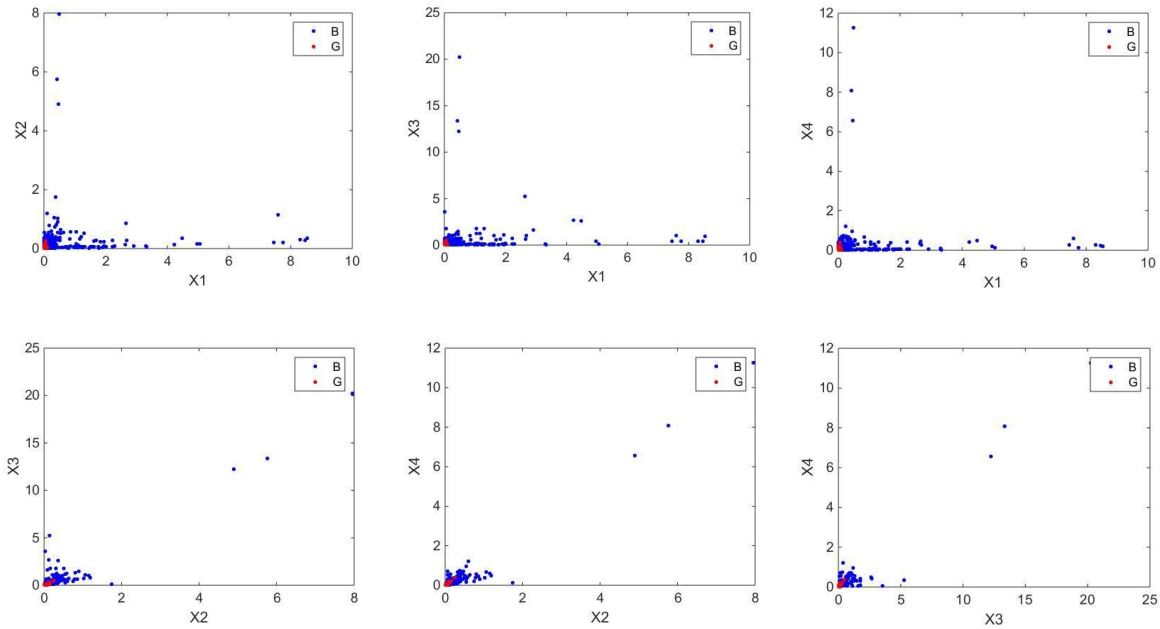


Figure 2 – Scatter plots by pair of variables. It has been identified the target value for each point of the scatter plot by using different colors: red for accepted lens (Good) and blue for the rejected ones (Bad).

An analogous analysis in two dimensions has been conducted for a reduced sample in which the furthest data from the centroid – point estimated taking the mean as the value for all the variables - have been removed. For this analysis, it has been chosen the euclidean distance as the distance metric and the percentile 95 of the distances to the centroid as the threshold for excluding the points for the following plots.

Some remarkable outcomes on the basis of the observation of these plots are related to the location of the points in accordance with the target value. Unlike the cases of rejected lenses, the ones associated to the accepted lenses show a high level of concentration around the origin in all the plots. However, the regions at these plots are not unmistakably defined for both target values, providing a common area between both regions, in which there is a majority of rejected lenses cases but also some cases of accepted lenses.

Furthermore, there is a well-defined linear and not negative, if not homoscedastic, relationship between pairs of all the variables but the first one with the reduced sample.

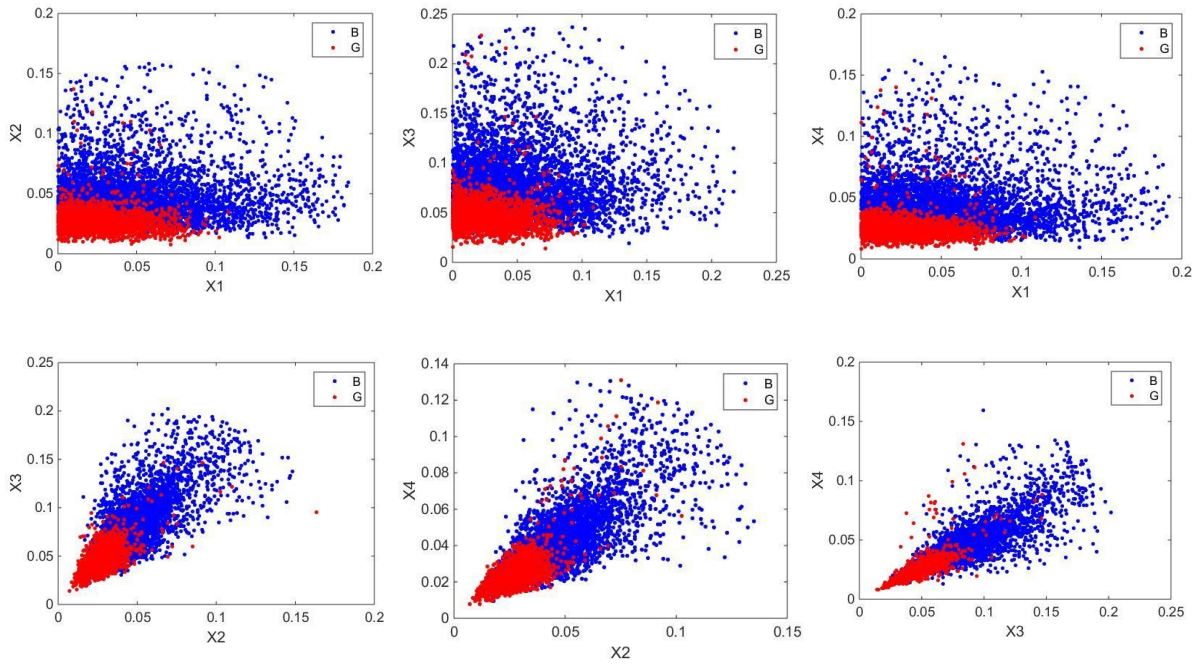


Figure 3 – Scatter plots by pair of variables (reduced sample). The furthest values from the centroid (far over the percentile 95 of the - euclidean – distance to the centroid) have been removed. It has been identified the target value for each point of the scatter plot by using different colors: red for accepted lens (Good) and blue for the rejected ones (Bad).

As none of the individual variables or pair of them seems to provide enough information to classify the lenses correctly, it has been decided to work initially with the four variables. More over, the methods designed or selected for these four variables have been also applied for the extensive database comprised by more than one thousand variables.

At this point, and for the purpose of performing an aggregated analysis with all the four variables, heatmaps were represented on the basis of various distance matrices among the points in R^4 defined by the standardized data of the overall sample, excepting target. For ease of interpretation the sample was firstly ordered by the target value.

For a classification problem, an ideal heatmap would present two opposing colors. The intra-group section would have the color reflecting a large closeness among pairs of cases with identical assessment of the target, whilst the contrary color, for large distances, would appear in the inter-group section.

As Figure 4 shows for the euclidean distance matrix, and as already observed for two variables in the scatter plots, heat-maps have revealed an apparently higher concentration among the cases of accepted lenses, but not the other desired properties: proximity of the bad cases with respect to each other and large distance between any two cases with different target values.

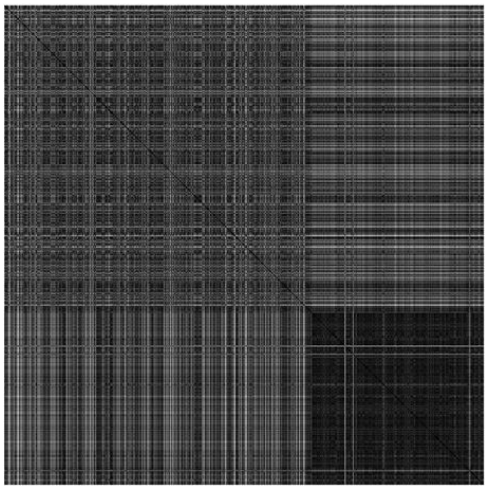


Figure 4 – Heatmap from standardised variables values by using an euclidean distance matrix. The graph represents the bad cases at the top and on the left section whilst the good cases are located at the bottom and on the right of the figure. Color grading used goes from white (separation of two points) to black (proximity).

3 .- TECHNIQUES

The information extracted from the data exploration suggests a distance-based approach should be used. Visualization of features data show us a physical separation, in the space of features, can been done. Because of these reasons, several techniques based on distances have been tested. Abording the problem of classification, these 4 classifiers have been constructed:

- Centroid method
- Shortest average distance
- K-Nearest Neighbors
- Support Vector Machine

For all the classifiers euclidean distance is considered. Classifiers have been constructed by using MATLAB.

3.1.- Centroid Method

The idea of this technique comes from figure 4. That led us to know that the samples of one of the labels are distributed in a more compacted way than them of the other label. That makes us think that we can define a hypersphere in the space of the features, whose center is the centroid of the samples related with that label.

A new sample to classify is represented in the space of the features. If it is inside the hypersphere, it is classified with that label. If not, with the opposite label. For testing, cross validation with four subsets have been used. The same in the other classifiers.

The radius has been determined by the computational experiments, looking for values near the distance between the centroids of that samples defined by both labels. Taking different values of the radius, a ROC curve can be drawn, in order to compare with others techniques.

3.2.- Shortest average distance

When all the training samples are represented in the space of the features, another classification by taking into account the two labels' samples can be used. When a new sample is wanted to classify, all the distance between it and the samples of one label are measured and stored. Same with the other label.

Then, the 20th percentile of the nearest distance of the smallest group (samples are not equally distributed between good and bad lenses) are chosen as a threshold. Average of the distances less than that threshold for one sample is calculated, and the same task for the other label. Whatever is the shorter distance, is the label the classifier assigns.

This value of 20th percentile works a threshold, as in the first classifier. This implies, also, a ROC curve can be constructed varying the value of the percentile.

3.3.- KNN

The third method tried is the well known K-nearest neighbors. In the same space of features as the other methods, training instances and a new sample are represented. The K nearest instances to the new sample are considered to classify.

Label assigned depends on the popular label of the K neighbors. If K=1, the classifier puts the same label as the nearest training instance. Several K values have been tried, and with K=1 it has been assumed that the accuracy reached is good enough.

This is the only classifier that, due to its implementation, just have been taught with the data set of 4 features.

3.4.- SVM

The last classifier tried is also based on distances. With all the training instances represented, the SVM classifier tried to define a hyperplane in space that delimit the samples of one label and the samples of the other. Linear SVM has been tried in this work, which means an hyperplane in the space of features.

Usually, not all the data can be strictly delimited by this construction. That is the reason because the SVM is a task of optimization. Besides, not only correct classification is searched, but also maximization of the distance from the samples to the hyperplane.

4.- DISCUSSIONS AND COMPARISONS

The above-described algorithms have been compared with each other on the basis of their performance of classification. This feature has been assessed by using a standard statistical measure, the percentage of true value or accuracy. The choice of that indicator rather than the other evaluation criteria derived from the confusion matrix, as the true positive rate – TPR or recall- and the true negative rate –TNR or specificity-, has been driven by the company's own valuation bases, according to which the same penalty for classification failures would apply to both target values.

The confusion matrix has been calculated by using a 4-cross fold validation. This technique consists of partitioning the original sample into 4 subsamples and replicating the process of model training and testing four times by using in each cycle one of the four subsamples as test set and the three remaining ones for training. The resulting confusion matrix has been determined by averaging the four confusion matrices obtained running the model on the test set in the various cycles.

Following table presents a summary of the results of accuracy, TPR and TNR achieved for the four algorithms analysed on one or both datasets available. As can be seen, all accuracy values are quite similar, within a very narrow range between 85% and 88%, despite the wider variation observed in the other complementary validation criteria. So, there is no a clear evidence for choosing one algorithm vs the other three ones by basing only on the accuracy measure.

Method	Data set	Accuracy	TPR	TNR
Centroid distance	4f/1058f	86%	79%	89%
Shortest average distance	4f/1058f	85%	86%	82%
Quadratic SVM	1058f	88%	90%	83%
KNN (K=1)	4f	88%	92%	81%

Table 1– **Summary table of methods considered.** It includes information about the dataset which it has been applied on and about the assessment of its classification performance.

Focusing on TPR and TNR (see Table 1) , it can be noticed the existing trade-off between both measures and also the wider range for TPR, a sign of the imbalanced data. Under this condition, the learning methods trends to be biased towards the majority group and the accuracy measure could not be a good metric for classification performance. So, it should have been aware of this characteristic of the data whenever the failure cost would be different for each group, being the minority one the most valuable concept to be learned, but also whenever we are trying to train the model homogeneously from both classes.

The ROC curve is a chart for summarize the classification performance in which the pairs (TPR, 1-TNR) are represented for different discriminants, allowing a choice of the best parameter values according to the classification performance.

Following figure shows a comparative of the ROC curves resulting for the classifiers based on the centroid distance and the one based on the shortest average distance. It can be seen that the two curves are quite similar. This suggests, in those two classifiers, threshold works similarly.

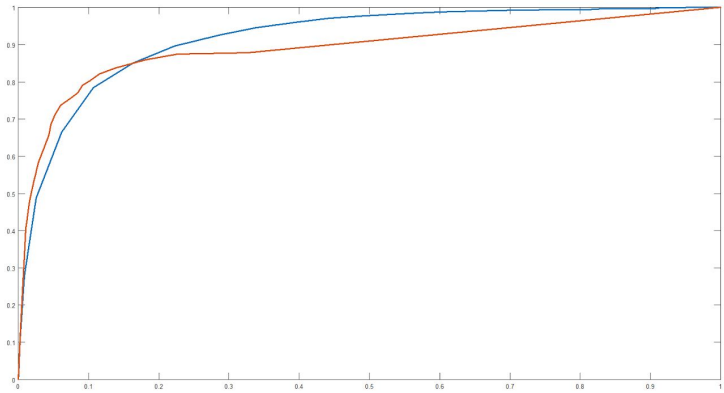


Figure 5 – **ROC curve**. Centroid method (blue) vs shortest average distance method (orange)

5.- CONCLUSIONS

Data exploration has been revealed as a good point to begin with. Showing us how data is distributed led us to try some types of classifiers. This is because the smaller class presents its samples nearer each other than in the other class. All of that using euclidean distance, a classic metric of distance.

Other information which can be extracted from data exploration is that deleting outliers might be, in practice, almost useless. A sample can be “outlier” looking at some features, but not looking to others. This problematic situation can be solved using classifiers that don’t take into account outliers, or reduce the influence of these. That is the idea behind centroid, shortest average distance and KNN classifiers.

What is more, looking at the accuracy and ROC curves, not a significant difference can be found between using 4 or 1058 features with the classifiers tried. Actually, there is neither contrast among the different classifiers. It might suggest distance classifiers are good, but have an accuracy limit around 85-88%

So, with these distance classifiers, working with all the features can be considered as a waste of memory space and CPU time. In other words, with these distance classifiers based on euclidean distance, because not a high improvement can be found with 1058 features instead of 4, there is no point in storing and processing all the data.

6.- FUTURE WORKS

The methods applied are ultimately based on a distance metric. In all the classifiers the distance used has been the euclidean distance due to the heatmap by using this metric. However, those classifiers, and other ones, might obtain better accuracy by using other distances, as Mahalanobis distance, that adds the variance-covariance matrix in its formula.

However, the good results of the heatmap don't imply that other standard learning methods for supervised classification couldn't be analysed for this problem. Alternative methods that we propose are those which treat the data as images.

This implies working with the full set of original features as pixels, without the information loss resulting from considering the sample as a tuple of unstructured variables. Indeed, a sample, represented as a vector, is a structured information; to keep all the information collected at each sample, it should be treated as an image. Instead of the standard methods that could work for general approach, such as decision trees or simple neural networks, it could be mentioned classifiers of images such as convolutional neural network (CNN).

All of the above-mentioned methods, applied and alternatives ones, classify each case within an unique class. However, it can also be considered that a case comprises a part of both classes, providing the fuzzy logic the percentages of membership in each one.