Introduction to Machine Learning and Data Mining

Jullian Bellino - Simon Thordal

September 29, 2014

Abstract

This report is the first part of the project given to students in "Introduction to machine learning and data mining" DTU's course. The goal of this part is to be able to visualize and understand a specific dataset. This one is composed of informations about several types of glass.

1 Visualization

1.1 Introduction

The data set is about classification of glass found at crime scenes in one out of 7 different categories. We found the data on the Machine Learning Repository. The classification is based on physical and chemical attributes such as refractive index and weight percentage of different chemical elements. An example can be seen in table 1.This is part of the training data and as such it includes the type each sample, however in the test dataset this information is not included, since determining the type is the goal of the data. Note that the dataset includes no

ID	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.00	0.00	1

Table 1: Example of attributs of data.

examples of type 4, so this category is excluded in the report. In addition we don't have any missing attribute values.

Classification & clustering

As an initial guess it would make sense to treat the data and the task as a classification problem, where clustering can be used to determine the type of glass based on the given attributes. It means we can determine type of the glass if this one has some similar values with an existing group. This approach could be chosen since at least some types can be determined based on the performed PCA analysis using the 1st and 2nd componenent.

Regression

We don't really know if it makes sense to use the regression in this case. Indeed our goal is to recognize the type of glass, in order to get some answers and maybe help the investigation in the case of crime scene. So we always have a group of data, and we'll often compare these one with old groups already analyzed. It's difficult to find a logic (it means an equation according to the regression model) in order to predict next data.

Association mining

In the same idea of previous tools, it would make sense to try to predict the type of our data. With the association mining it would be possible de determine some rules between attributes in order to determine which type of glass are the data. For example if a high presence of some specific attributes determines a glass of type X, we are able to predict which kind of glass are for a part of data.

Anomaly dectection problem

This kind of method can be apply in the case of a crime scene. It can be useful to identify which type of glass are really interesting, and if they were not here before the crime. It's in order to keep just interesting values and remove all outliers.

Actually the classification using clustering seems to be the method that makes the more sense for our problem. We mean, how to recognize a type of glass according to the attributes we have.

1.2 Presentation of data

The data and its summary statistics can be seen in table 2. Note that even the type is included here although it is not an attribute in itself, but the goal of the exercise.
Since the means have a difference of up to three orders of magnitude it will be hard to visually compare data in a single plot, so it makes sense to standardize the data before performing the PCA.

Name	Attribute type	Discrete / Continuous	Mean	σ^2	Range
RI	Ratio	Continous	1.5184	0.0030	0.0228
Na	Ratio	Continous	13.4068	0.8184	6.6500
Mg	Ratio	Continous	2.6761	1.4405	3.9800
Al	Ratio	Continous	1.4465	0.4999	3.2100
Si	Ratio	Continous	72.6550	0.7741	5.6000
K	Ratio	Continous	0.4991	0.6530	6.2100
Ca	Ratio	Continous	8.9579	1.4264	10.7600
Ba	Ratio	Continous	0.1759	0.4982	3.1500
Fe	Ratio	Continous	0.0573	0.0976	0.5100
Type	Ratio	Discrete	NA	NA	NA

Table 2: Summary statistics for the attributes.

1.3 Analysis of normality

Initially we thought about analyse the normality of our data. Indeed the normal distribution is the most common distribution which is used in life's phenomenons. That's why you can see below the histogram of each data, in order to notice what kind of distribution it could be. Moreover with the idea of find a normal distribution, we also plot the normal curve on each histogram (see figure 1).

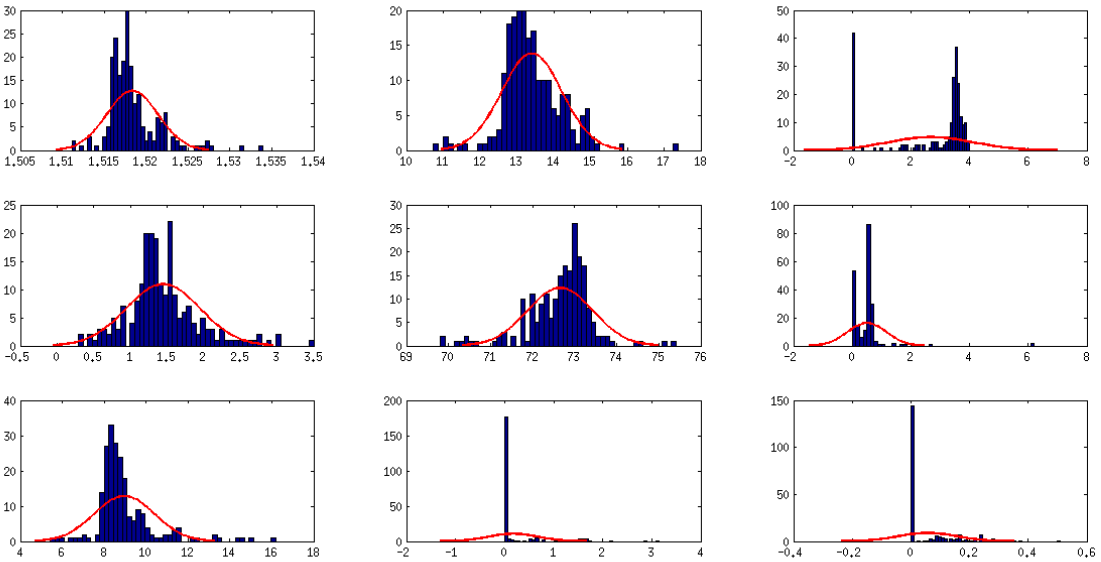

Figure 1: Histograms for each data attribute matching with normal curve.

In this case we can see that none exactly follow a normal distribution. Possible reasons could be that we don't have enough data to see the distribution, we have too many extreme values, too much noise or too many outliers. Alternatively the distribution of data just is not normal and we have to work with non-normal data. A Kolmogorov-Smirnov hypothesis test was also performed on each dataset to test for normality and it was always rejected (with 5% of error). That's why we began to find an other kind of distribution for these data.

These distributions were found using Matlabs built in tools (see table 3). But we have to notice that we have several cases. Firstly the closest from normal distribution is the location-scale distribution. It's a model like the normal model, however it's more adapted for extrem values, when data don't exactly fit the normal distribution. That means some extreme values can explain why we can't fit a normal distribution.

Attribute name	Distribution type
RI	Location-scale
Na	Log-logistic
Mg	Generalized Pareto
Al	Log-Logistic
Si	Location-scale
K	Exponential
Ca	Location-scale
Ba	Generalized Pareto
Fe	Generalized Pareto

Table 3: Results of distribution's analysis.

Then we recognize an other distribution : log-logistic. This is a kind of distribution which is not so far from the normal distribution. The only difference is about tails which are heavier. One more time we find a distribution similar from the normal. That's why we can explain it by not enough values or too much extrem values.

The two others distributions found are : Pareto and exponential. These models seems to fit the last three attributes (we'll see in the PCA that these attributes are less important than the others and why). It's a bit strange to fit these in our case where attributes are for the most part, chemical elements. We can guess it's maybe because we don't have enough data to be sure that they are normal or not. But in our case is not a real problem because we'll focus on the first five attributes which are more or less close from a normal distribution.

All in all however it should be noted that this might simply be a case of malformed data or too small datasets and that it is very possible that we would get different results from a larger sample.

1.4 Principal Component Analysis

Principal Component Analysis is the process of finding a basis for a data set, that explains the most of the variance of the data. This is useful as the components are compounds of the original attributes, meaning a specific organization or combination of components can be used to make a classification of the data and less important components can simply be discarded from the analysis, leading to dimensional reduction of the data set.
The PCA is performed on data that has been standardized by subtracting the mean and dividing with the standard deviation. The parameter ρ tells us the proportion of the variance that is explained by each of the components, a plot of which can be seen in figure 2. This can be used to determine which components

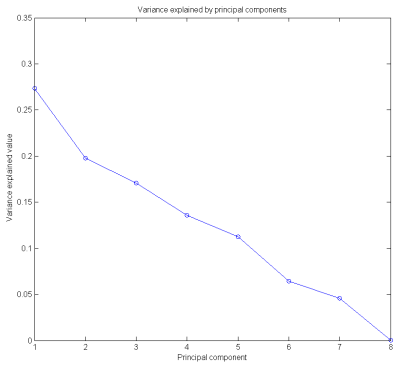

Figure 2: Variance explained by each principal component

should be kept and which could be discarded, which is the dimensionality reduction part of principal component analysis. In this case the first five components explain 88% of the variance, so the dataset can be explained by those while excluding the remaining three components, without losing too much information about the data.

The matrix V can be used to show the loadings of each of the original attributes in relation to the principal components. It can be seen from figure 3 that the first and second component respectively place a high emphasis on content of Mg and Ca versus K and Si. Note that the signs of the loadings is arbitrary, so it would have been possible to use absolute values instead. The data projected onto the principal components can bow be presented, which makes the most sense if paired with a loading plot showing the loading of each of the original attributes on the principal component. Although five components have been kept only the plot for the first and second principal component is shown here at is has by far the most explanatory power and is fairly easily interpreted.

From the score plot of principal component one and two in figure 4 it can be seen that glass from headlamps can be easily identified by the calcium, silicon and magnesium content whereas containers can be found by the second principal component, due to the potassium content, however they are not easily distiguished from building windows and tableware. Using clustering it would

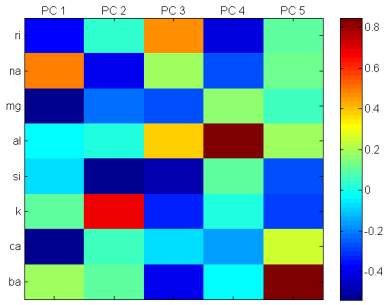

Figure 3: Components vs attributes

be possible to identify the headlamps, however the other attributes might require some kind of supervised learning technique to be properly identified. It is also possible that iteratively using clustering over each principal component could yield good results, i.e. first removing identifying as much as possible using the first two components, classifying them and removing them from the set, then try again with the next principal component and checking if removing the already categorized results makes a new cluster appear and so on for all the components.This is something we will check in the next report.

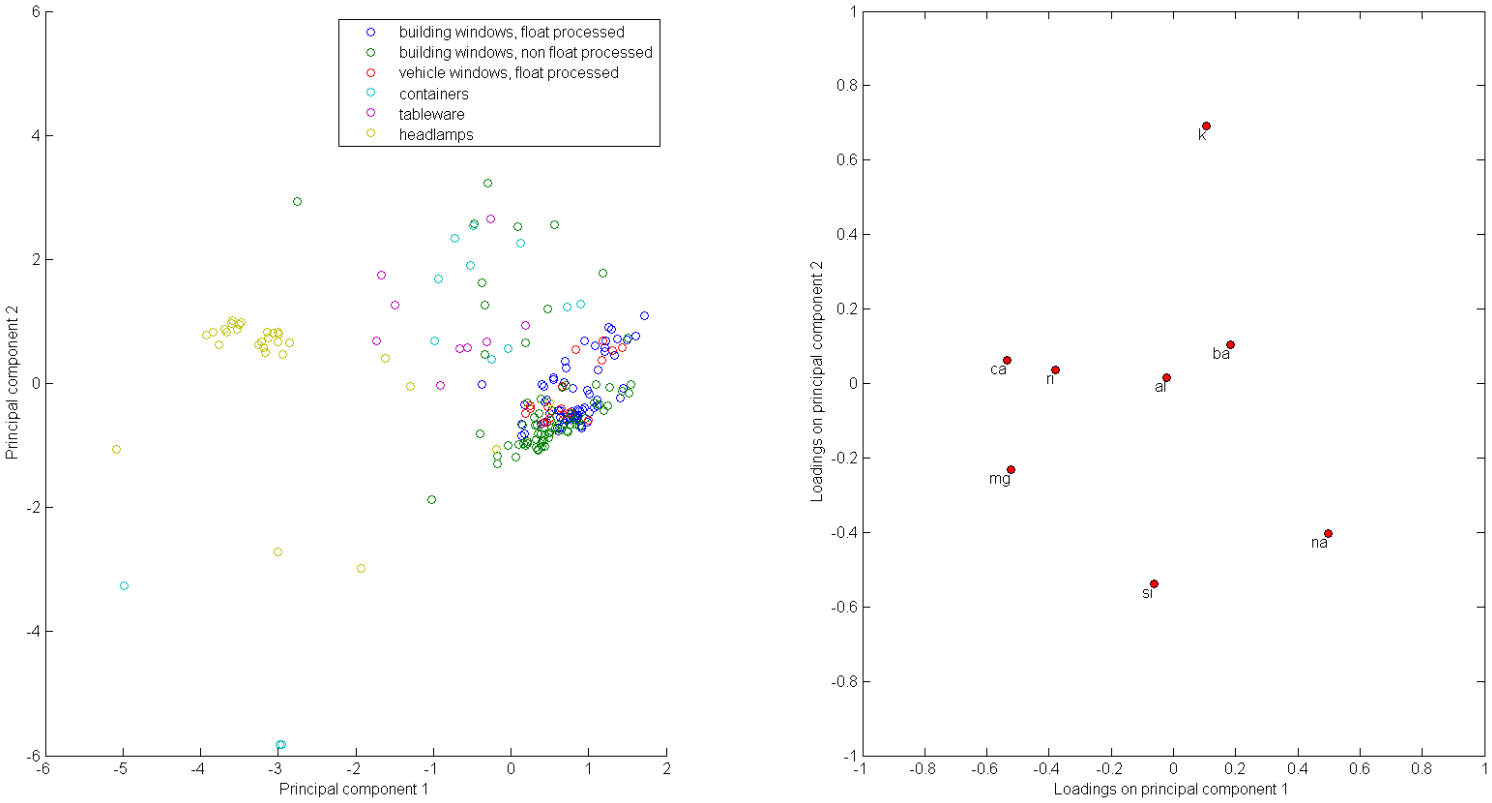

Figure 4: Score plot for PC1 and PC2 along with the loading plot for the components

1.5 Conclusion

To conclude, thanks to this analysis we are now able to understand what each attribute means and using PCA we can determine the usefulness of each attribute or more correctly the components. From the analysis of normality we have presented some doubt about the distribution of our data, but hopefully we will still be able to use the techniques we're planning despite of this and the PCA also seems to suggest that clustering would be a good first choice for this particular data set. Hopefully this will allow us to make a good classification and achieve the goal of determining the type of data from the given chemicla concentrations.