Data Visualization

Classification of credit institutions and visualization of data

by Amine Boumaaz & Mohamed Zaki Chellali

12 January 2017

Abstract : *As part of the cyclical monitoring of the consumer credit market and in order to distinguish business practices from conventional banking networks from those offered by consumer credit institutions generally backed by large retailers or institutions on the retail sector, it is essential to propose a classification of establishments in order to distinguish between these different types of actors. For this, we use unsupervised learning algorithms, which allow to answer to this kind of problem. If the results of the Hierarchical Ascending Classification are visualized by using the hierarchical tree, it remains relative to this method, and it is not possible to do the same for other algorithms. We propose in this project methods of visualization of clustering algorithms, notably thanks to the method of parallel coordinates, and radar-chart. We also introduce a new technique we didn't used, but that we find interesting because it that does both clustering and simplify visualization in the same time.*

Introduction

The financial crisis has highlighted the need to have rapidly available indicators of financial health of the banking market, reviving the interest in banking economy in the methods of classification of credit institutions. Indeed, these methods facilitate market analyzes by allowing groupings of financial entities sharing structural characteristics ; A segmentation of the banking market makes it possible not only to give a synthetic view of the sector, but also to concentrate the analyzes on a particular type of actors, in particular establishments granting consumer credit.

In addition to the variety of types of credit referred to as consumer credit, these credits have the distinction of being distributed not only by general establishments, such as those belonging to large banking networks, but also by specialized establishments, frequently backed by mass retail chains or car manufacturers.
The types of credit, the amounts granted but also the rates applied differ according to the type of lending institutions.
The Bank of France currently publishes quarterly interest rates by type of lending institutions. The data used are those of the Bank of France, and are of a confidential nature.

The variables used are those relating to bank accounting. Therefore, we will use the logarithm of the size of the balance sheet, the shares of different types of credits (consumer credit, leasing, home loans, credit to non-financial corporations) in relation to total credit granted, inter bank liabilities, the share of assets.
The banks are normalized, in order to have the same scale.

As a first step, the objective is to classify credit institutions by applying unsupervised learning methods to bank account data.
From this classification, the second step aims at visualizing the results of the clusters obtained.
We also present a clustering method, which allows both to classify banking institutions,and visualization of the clusters.

Statistical Background

Two main families of exploratory data methods are generally cited in the literature, size reduction methods and classification methods. Unsupervised classification methods, or unsupervised learning methods, or clustering methods, have been developed to partition a group of individuals into several subgroups from a set of observations and variables.
Unlike supervised learning methods, partitioning is

performed without a prior information on the data structure. The individuals belonging to the same subgroup form a homogeneous class according to the variables considered, while having profiles quite different from the individuals of the other classes.
The objective is to minimize the intra-class variance while maximizing the inter-class variance, the two objectives being potentially competing. Several categories of unsupervised learning methods exist : a distinction between probabilistic approaches and geometric approaches is generally proposed.
We will limit ourselves here to the geometric approaches, as well as to the visual potential of each method.

K-means

This method is a classic of unsupervised learning methods and its scope extends from day to day. Whether it is to establish a segmentation of a population in a marketing context, to identify the patterns making up an image in medical imaging (recognition of a tumor on an x-ray), or for satellite image recognition, it didn't ceased to prove itself by helping to interpret the results. This method has the advantage, unlike the Hierarchical Ascending Classification, of being less costly in computing time.

Having some individuals $(x_1, x_2, ..., x_n)$ of length n, those methods aims at partitioning individuals into $k \leq n$ partitions $S = (S_1, S_2, ...S_k)$, by minimizing the distance between the individuals in the same cluster :

$$argmin_S \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \tag{1}$$

where μ_i are the means in each cluster.

We start by randomly assigning (or defining if we have a prior information) k points in space, which represent the centers of the classes. At each iteration, the individual closest to one of the points, according to a predefined distance measure, is assigned to this class and the mean or center of the two points becomes the new center of the class. We recommence the operation until the individuals no longer change class. The number of possible partitions is k, and at each step of the algorithm, the cost function decreases strictly, ensuring the convergence of the algorithm. However, the solution obtained results from local and non-global optimization.

If this solution has the advantage of being more efficient in term of computation time than the hierarchical methods, it nevertheless requires to know the number of classes k to be used. Moreover, contrary to the CAH, the visualization of the results is more delicate because there is no dendrogram.
Generally, to visualize the clusters obtained, we use some methods of reduction of dimensions such as PCA, which allow us to plot the results on two axes only.

Self-Organizing Maps :

Unsupervised learning also includes these methods of reducing the size of the starting space. These methods differ from conventional clustering methods and are not intended to classify individuals, but rather to provide a synthesis of information. However, there are also methods of clustering conciliating reduction of dimension and classification of individuals

The objective of the SOM is to reduce the size of the starting space, while preserving not the distance between individuals, but rather their neighborhoods. This method is useful when the number of variables is considerable, because the data is sparse in space. It is then difficult to identify clusters of significant size if the population is composed of few individuals, and visualize them correctly.
The SOM differs from the methods seen previously, in the sense that we build beforehand a grid of fixed dimension, constituted by vectors, called neurons. The shape of the chosen map (rectangular, hexagonal, or uni-dimensional) determines the number of neighbors of each neuron. The objective is then to bring neurons closer to individuals so that they represent them at best.

If R^P represents the starting space consisting of p variable and K the total number of neurons on the map, each neuron k is associated with a reference vector w_k of this space.
At each iteration $t \in T$, we start by projecting an individual x_t chosen randomly on the grid. Then, we determine the winning neuron c satisfying :

$$d(w_c(t), x(t)) = \min_{k \in [|1,K|]} (d(w_k(t), x(t)) \tag{2}$$

This neuron, but also its neighbors, will be modified to more closely resemble the projected individual. For this purpose, the reference vectors of these neurons are modified according to the formula :

$$w_k(t+1) = w_k(t) + \alpha(t) h_{ck}(t)[x(t) - w_k(t)] \tag{3}$$

The parameter $\alpha(t)$ represents the learning step. The higher it is, the more neurons will be modified to correspond more closely to the projected individuals. One way to choose this parameter is to make it decrease linearly according to the number of iterations.

The function $h_{ck}(t)$ defines the correction performed on the neurons located near the winning neuron c. We usually use the Gaussian function :

$$h_{ck}(t) = exp(-\frac{||r_c - r_k||}{2\sigma(t)^2}) \qquad (4)$$

Where r_c, r_k are the coordinates of the two neurons on the grid, and $sigma(t)$ is the neighborhood radius of the neuron c in the iteration t. According to this function, the amplitude of the adjustment depends on the distance of each neuron to the winning neuron. Thus only the reference vectors of the neurons close to the winning neuron will be adjusted.

At the end of the iterations, each observation is attributed to its victorious neuron. Two neighboring neurons on the map will represent close observations in the starting space. The conservation of the neighborhood is thus assured. We then obtain the projection of each variable on the map :

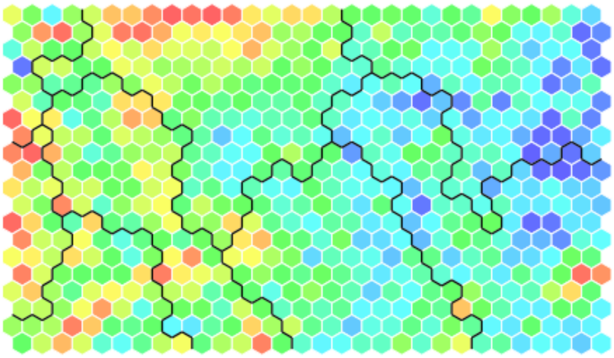

FIGURE 1 – Self Organizing map

Related Works

The application of unsupervised learning methods on financial data and the visualization of results has already been done by several bodies, notably the ECB.

The topic of European financial integration has been at the forefront of economic research in recent years, in particular sparked by the advent of Economic and Monetary Union and the endeavours to create a Single Market for Financial Services.

Cluster analysis is a useful tool to examine complex relations among national characteristics and international linkages without imposing any a priori restrictions on the interrelationships.

In order to visualize the clusters obtained, the average of each group of the different variables used is computed, and showed in spider plots.

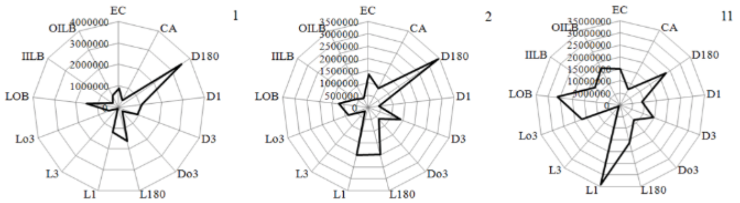

FIGURE 2 – Mean values for clusters

Other works, still by the ECB were made due to the recent global financial crisis that has demonstrated the importance of understanding sources of domestic and global vulnerabilities, that may lead to a systemic financial crisis.

Early identification of sources of vulnerability is important as it would allow introduction of policy actions to decrease further build up of vulnerabilities or enhance the shock absorption capacity of the financial system.

Dimensionality of the problem complicates visualization, since a large number of indicators are often required to accurately assess vulnerabilities to a financial crisis.

In addition to the limitation of standard two- and three-dimensional visualizations in describing higher dimensions, there are challenges of including a temporal or cross-sectional dimension.

The Self Organizing Map (SOM) were used, because they combine the aims of projection on lower dimensions and clustering techniques.

It can provide an easily interpretable non-linear description of the multidimensional data distribution on a two-dimensional plane without losing sight of individual indicators. Thus, the two-dimensional output of the SOM makes it particularly useful for static visualizations, or summarizations, of large amounts of information.

Visualization

For the visualization of our results and multidimensional data in general, it is possible to use a dimension reduction method in order to retain only the most important axes, such as principal component analysis.

In our case, the inertia retained by the first 2 axes reaches 45% of the information, which is not sufficiently representative to be limited to these two axes. Moreover, the direct interpretation of the axes of a PCA is much more complicated. Therefore, we will use another method that allows to visualize multidimensional data : the parallel coordinates.

Parallel coordinates

A parallel coordinate plot maps each row in the data table as a line, or profile. Each attribute of a row is represented by a point on the line. This makes parallel coordinate plots similar in appearance to line charts, but the way data is translated into a plot is substantially different.
For each bank type, it is now possible to plot a profile of how the different parameters are distributed.
The experts can now see which bank types are similar to each other in loans distribution, by comparing the profiles of each banks to each other. This is where the parallel coordinate plot is really useful, to compare profiles in order to find similarities.

Adding more dimensions in parallel coordinates involves adding more axes. The value of parallel coordinates is that certain geometrical properties in high dimensions transform into easily seen 2 dimension pattern.
When used for statistical data visualization, three parameters becomes important : the order, the rotation, and the scaling of the axes.

- The order of the axes is critical for finding features, and in typical data analysis many reorderings will need to be tried.
- The rotation of the axes is a translation in the parallel coordinates and if the lines intersected outside the parallel axes it can be translated between them by rotations.
- The values in a parallel coordinate plot are always normalized. This means that for each point along the X-axis, the lowest value in the corresponding column is set to 0% and the highest value in that column is set to 100% along the Y-axis.

It is also possible to brush data, to interact with graphs in figure windows in which it is possible to drag a selection rectangle around data points to highlight observations. Highlighting makes it easier to focus solely on the chosen cluster, or individuals who meet certain conditions.

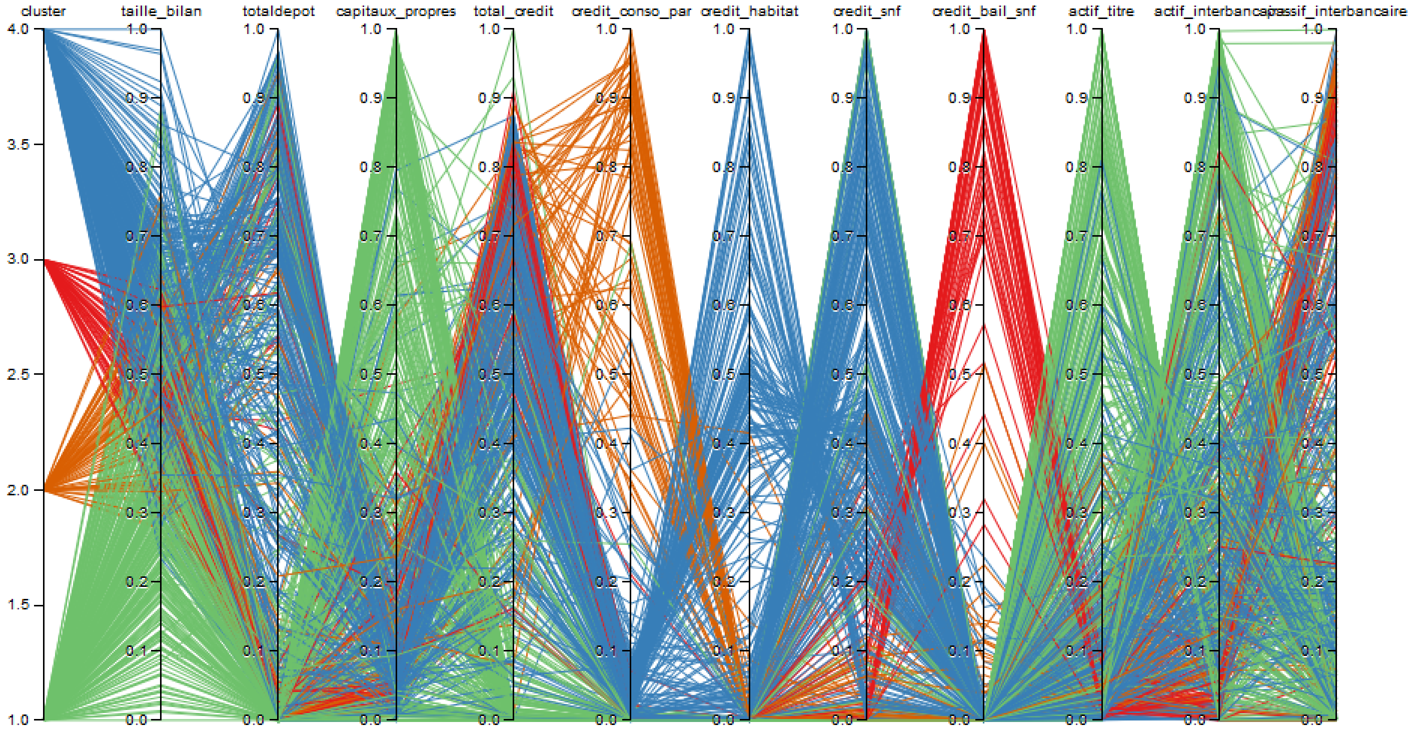

Figure 3 – Parallel coordinates of clusters

Radar Chart

A radar chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point.
It consists of a sequence of equi-angular spokes, with each spoke representing the mean of a cluster on one of the variables.
The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each spoke. This graph is used to answer to two questions :

— Are there differences between the clusters

— What are the discriminant variables

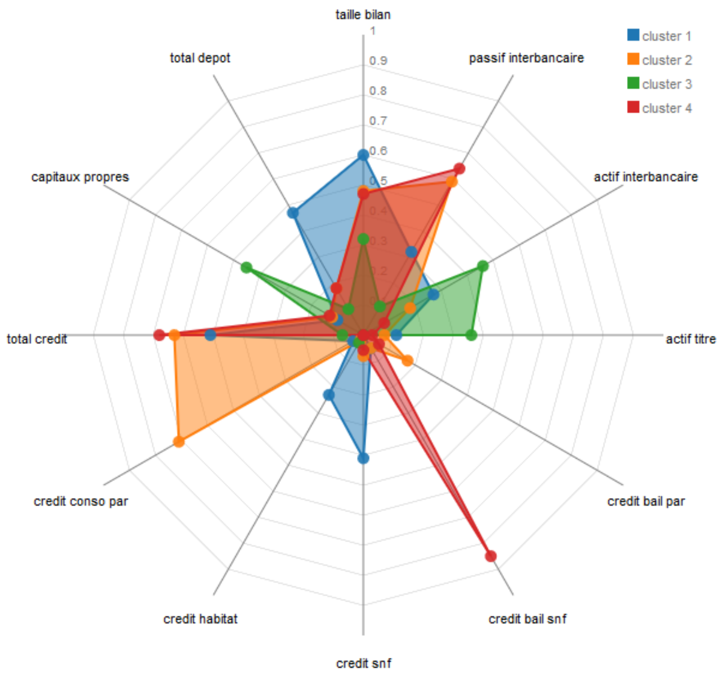

FIGURE 4 – Mean values for clusters

Results interpretation

We discover four types of banking institutions :

- Investment banks, which are characterized by a higher ratio of equity than other banks, because they operate on the financial markets, in order to reassure future investors.They are also lenders in the banking market, and have a fairly high asset value.
- General banks, which are financed mainly through the deposit of individuals, hence the high value of this variable, and which are fairly stable on the inter-bank market. They also have credit activities, notably for non-financial corporations, but also for lending to housing.
- Consumer credit banks, which constitute our target cluster and represent banks providing credit to households, whether in the form of restricted credit, personal loans or revolving credit. They are financed particularly on the interbank market.
- Credit banks lease to non-financial corporations, which share the same structural characteristics as consumer credit banks except for the types of credit granted.

Discussion

The methods we used to visualize the clusters obtained are not always well suited for those kind of problem far from it.

In parallel coordinates, each axis can have at most two neighboring axes (one on the left, and one on the right). For a p-dimensional data set, at most $p - 1$ relationships can be shown at a time.
In time series visualization, there exists a natural predecessor and successor ; therefore in this special case, there exists a preferred arrangement but it is not our case.
However when the axes do not have a unique order, finding a good axis arrangement requires the use of heuristics and experimentation. In order to explore more complex relationships, axes must be reordered.

The spider plot also have their limit. It is impossible to know the number of individuals making up each cluster. Moreover, the use of the mean for each cluster can be problematic, in the sense that the outliers can influence the results. And why should we use the average rather than another statistical indicator ?

Conclusion

In order to highlight the clusters obtained by the unsupervised learning method, we used mainly methods such as radar chart, or coordinate parallels. However, these methods quickly find their limits, especially when the dimensionality of the data increases. The use of self-organizing maps can overcome this problem, since a two-dimensional map is used to visualize multi-dimensional data, and this method also has the advantage of being a good clustering method.

Références

[1] P. Sarlin, T. A. Peltonen, "Mapping of the financial stability", Journal of International Financial Markets, Institutions and Money. october 2013

[2] G. Cabanes, Y. Bennani, "Un algorithme de classification topographique non supervisée à deux niveaux simultanés. 8èmes journées d'Extraction et Gestion des Connaissances (EGC'08), pp. 619–630, Sophia-Antipolis, France (2008)".

[3] "Find Content Gaps Using Radar Charts". Content Strategy Workshops. March 3, 2015.

[4] N. Dardac, I. Boitan, "A cluster analysis approach for bank's risk profile".

[5] G. Halaj, D. ZOchowski, "Strategic Groups and banks performance".

[6] T. Jagric, T. Markovic-Hribernik, "The integration of the European financial sector – the case of the banking sector".

[7] Euro area banking sector integration using hierarchical cluster analysis techniques(2006), by Christoffer Kok Sorensen and Josep Maria Puigvert Gutiérrez

[8] Chantal Hajjar (2014), Cartes auto-organisatrices pour la classication de donnéses symboliques mixtes" (2014).

[9] V. Vagizova, K. Lurie, Ihor Ivasiv, "Clustering of Russian banks : business models of interaction of the banking sector and the real economy".

[10] R. Ayadi, E. Arbak, W. P.r de Groen, "Business models in European banking pre-and post crisis screening.

[11] T. Kohonen, E. Oja, O. Simula, J. Kangas, A. Visa , "Engineering applications of the self-organizing map" (1996).