For Visualization-Based Analysis Tools in Knowledge Discovery Process: A Multilayer Perceptron versus Principal Components Analysis: A Comparative Study

Xavier Polanco, Claire François, Mohamed Aly Ould Louly

Programme de Recherche en Infométrie (PRI) - Informetric Research Program
Institut de l'Information Scientifique et Technique (INIST)
Centre National de la Recherche Scientifique (CNRS)
2 allée du Parc de Brabois - 54514 Vandœuvre-lès-Nancy Cedex, France
claire.francois@inist.fr, polanco@inist.fr

Abstract. Mapping knowledge structures is a key task in Knowledge Discovery in Databases (KDD). In order to display the thematic organization of knowledge, we compare and evaluate two different cartography approaches: principal components analysis (PCA) and a multilayer perceptron (MLP) in "self-association" mode. This kind of MLP can be used to perform a PCA when the activation function is set to the identity function. This allows us to look for the non-linear activation function which best fits the data structure. We present an evaluation criterion and the results and maps obtained with both methods. We notice that the MLP detects a non-linearity in the data structure that the PCA does not detect. However, the MLP does not express the non-linearity completely. Finally we show how a related component analysis (RCA), based on graph theory, provides representations of the inter-clusters relationships, compensating for the approximate nature of the maps, and improving their readability.

1. Introduction

We are concerned with the design and development of information analysis tools in which informetrics, computational linguistics and artificial intelligence techniques are combined. In particular, *informetrics* address the issue of applying metric or quantitative information analysis methods (i.e., statistics, probabilities and multivariate data analysis) to produce useful information (for a general statistical perspective on KDD, see [3]). The aim of our activity is to perform the analysis of information by computer using cluster analysis and cartography algorithms which represent the generated clusters in the form of maps. We apply this approach to the domain of scientific and technical information, i.e., publications and patents stored in databases (for details, see [7]; [10]; [11]; [9]).
In this article, we avoid using the terms *classification* and *classes*; rather we use *cluster analysis* or *clustering* and *clusters*, in order to distinguish at the conceptual

level two different techniques: a supervised method that classifies data into predefined classes or taxonomy, and a descriptive unsupervised technique that seeks to produce statistically some aggregations from data themselves. We use the terms *cartography* and *maps* to mean our statistical-based process of knowledge representation, that is *knowledge mapping*, as opposed to the knowledge representations and the processing of representations in logically-inspired ways (see, [8], ch. 6).

1.1 Process Overview

Clustering, cartography, and hypertext generation are the three main components of our approach. Informetric analysis of the information is divided into two phases: the first involves the generation of clusters using clustering procedures, in which learning is unsupervised (the user does not define classes), while the second consists of positioning the clusters on a global map in order to display the topical organization of knowledge. These two phases are data driven. A hypertext interface generator provides the user with a user-friendly interface displaying the global map, the topics or clusters and the documents themselves. The global map consists in an overview of the target documents set and then gives access to useful information organized by topics (clusters). This approach is implemented in the NEURODOC system.
The NEURODOC system uses the axial k-means method (AKM), i.e., an unsupervised winner-takes-all algorithm producing overlapping clusters, and a principal components analysis (PCA) for mapping. Our research is currently oriented to develop NEURODOC into a platform which can be used to apply artificial neural networks (ANNs) on bibliographic and textual data for clustering and mapping ([12]). Our interest in ANNs algorithms lies in the links which exist between multivariate data analysis and connectionist approach.
This article covers the cartography issue only, and concentrate on work concerned explicitly with mathematical means for comparing two different cartography techniques: PCA and a MLP. In the process of KDD, the maps are "visualization-based analysis tools". As Brachman and Anand ([2], p. 45) note: "The visualization produced is by itself a model, and the user can examine the visualization to determine its explanatory power (...) Appropriate display of data points and their relationships can give the analyst insight that is virtually impossible to get from looking at tables of output or simple summary statistics. In fact, for some tasks, appropriate visualization is the only thing needed to solve a problem or confirm a hypothesis, even though we do not usually think of picture-drawing as a kind of analysis".

1.2 Application Domain

The target data set used here for mapping knowledge structures was built for an unpublished study on "prion diseases" which was carried out in 1997 by PRI/INIST for the Life Science Department of the CNRS. The corpus consists of 1855 bibliographic records taken from the Science Citation Index CD-ROM (SCI). This

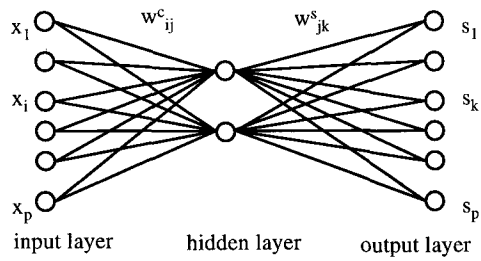
data set represents essentially the past five years of research published in 453 journals covered by the SCI. Documents (n objets) are indexed by 1750 keywords (p attributes) generated by an in-house linguistic engineering platform (ILC platform) applied on titles and abstracts. Using the AKM, the NEURODOC cluster analysis generates 30 clusters (m) described by a cluster matrix X ($m * p = 30 * 1750$) which is very multi-dimensional and sparse, that is essentially composed of null values.
In this paper, the clusters, that are obtained, are the data to be positioned on the global map. In documentary data analysis, the representativeness of the two axes of the map measured as an inertia percentage, is always very weak. In the case of the "prion diseases" corpus, the factorial plane of the PCA explains only 17.77 % of the inertia.. The MLP is of interest since it is able to detect non-linearities in the data structure and therefore produces a more representative cartography in theory.

2. A Non Linear MLP with One Hidden Layer

This section presents the MLP used for mapping the clusters obtained by the AKM. We do not use the MLP for its "associative properties" which perform a discriminant factorial analysis, but for its "spatial projection properties" which can be used to map the data. The learning is defined by the gradient back-propagation algorithm, which is both stochastic and incremental.

2.1. Network Architecture

The network architecture corresponds with the task that the MLP has to perform. Bourret *et al.* [1], continuing the work of Gallinari *et. al* [6] show that a linear MLP using the "identity" activation function ($f = Id$) can perform a PCA. To do this, the input and output layers have an equal number of neurons and the hidden layer has only two neurons. This MLP takes vectors of size p as input and compresses them in 2-dimensional vectors before restoring the information at the output. Once the convergence has been carried out, the outputs of the hidden neurons are used for a 2-dimensional map representation.
The data (clusters) are centered before to be passed through the MLP, implying that the network does not contain a polariser neuron, i.e., a neuron the output of which is always equal to 1, which would give as result a affine line. In this case, we know that the best-fit plane passes through the origin now situated at the center of gravity of the clusters.



For an input x, the output x^c_j of neuron j of the hidden layer is given by :

$$x^c_j = f\left(\sum_{i=1}^{p} w_{ij} x_i\right)$$

Similarly, the output s_k of neuron k of the output layer is given by :

$$s_k = f\left(\sum_{j=1}^{2} w_{jk} x^c_j\right)$$

Fig. 1. MLP architecture. Notation: l is the indice for the m clusters of the learning set; i, j and k are the indices for the first, second and third network layers respectively; X is the cluster matrix in which each row x is one of the input vectors, $x = [x_1, ..., x_i, ..., x_p]$; x^c is the output vector of the hidden layer, $x^c = [x^c_1, x^c_2]$ (in French "c" for "cachée" or hidden); s is the output vector of the network, $s = [s_1, ..., s_k, ..., s_p]$ (in French s for "sortie" or output); W^c is the weighting matrix of the hidden layer, each element of which denoted by w^c_{ij} represents the i^{th} weight of the j^{th} neuron of the hidden layer; W^s is the weighting matrix of the output layer, each element of which denoted by w^s_{jk} represents the j^{th} weight of the k^{th} neuron of the output layer; and lastly, f is the network neuron activation function

2.2. Learning Algorithm

The learning mode is a "self-association" supervised learning algorithm, that is to say that the network is trained to adjust its output as close as possible to the input. For each input, the individual squared error (*ISE*) corresponds to the Euclidean distance between the input and output vectors. The learning level of the network is known due to the mean squared error (*MSE*) which is calculated periodically. Its evolution has three phases: a relatively long plateau, a step-wise descent, and a final stabilization. The *MSE* is given by the following formula:

$$MSE = \frac{1}{m}\sum_{l=1}^{m} ISE(l) \text{ where for a data x : } ISE = \|s - x\|^2$$

The learning is stochastic, the network connections are corrected after processing each data x. Moreover, the standard gradient back-propagation algorithm is not incremental, and the connections of the two neurons of the hidden layer are calculated simultaneously. In order to perform a PCA, we must apply an incremental algorithm [5], and so once the connections of the first neuron have been calculated and have stabilized, the connections of the second neuron are then calculated. Once convergence has been achieved, the outputs of the two hidden neurons define the coordinates on the map for each cluster.

3. Statistical Assessment of Performance

This section presents the comparisons between PCA and a MLP using different activation functions. In attempting to perform this comparison, we present the PCA as a means of approximating a matrix by minimizing the *MSE*.

3.1. MSE definition for PCA

To place the clusters on a map (defined in the space R^2), the best result obtained using multivariate data analysis methods is the set of data projections on the first two factorial axes, also defined by the set of the coordinates of the first two principal components, the vectors c_1 and c_2. This corresponds to the second rank approximation of the cluster matrix X, that is $X^*(2)$:

$$X^*(2) = X\sum_{k=1}^{2} u_k u_k'$$

where u_k is the unit eigenvector of the matrix $X'X$ associated with the k^{th} largest eigenvalue λ_k, or the unit vector of the k^{th} factorial axis. This approximation is the result of the maximization of the squares of the projections of the m individual points (row vectors representing the clusters) onto the factorial plane created by the vectors u_1 and u_2, and which represents the visualization plane. This approximation also corresponds to the minimization of the sum of the squares of the distances of the m individual points from this factorial plane. But the sum of the squares of the distances corresponds to what is called the general squared error (*GSE*) in the field of ANNs. Once the PCA has been performed and the two principal components determined, we will be able to evaluate the *GSE* and then *MSE* is defined by:

$$MSE = \frac{1}{m} GSE = \frac{1}{m}\sum_{i=3}^{m} \lambda_i$$

where m is the number of clusters, λ_i is the i^{th} eigenvalue associated with the i^{th} factorial axis. This *MSE* corresponds therefore to the sum of the eigenvalues of the complementary space to the first factorial plane, which is defined by the two factors associated with the two largest eigenvalues. This *MSE* is the *minimum MSE* that the PCA can achieve, and is thus our means of evaluating the MLP.

3.2. PCA and linear and non-linear MLP

The MLP performs the approximation S_f for the cluster matrix $X(m,p)$. For the identity function $f = Id$, the network output becomes:

$$S_{Id} = Id\left(Id\left(XW^c\right)W^s\right) = XW^cW^s$$

The network must then find two weighting matrices W^c and W^s with sizes $(p,2)$ and $(2,p)$ respectively such that the matrix XW^cW^s is the best possible approximation of

the matrix X. This is equivalent to finding a matrix M of size (p,p) and rank 2 such that XM is the best approximation of X. This matrix is $X^*(2)$ (cf. § 3.1).
The weighting matrix $W^c W^s$ must therefore converge to this matrix M. In other words, the self-associative linear MLP (where $f = Id$) with two hidden units projects the data in the 2-dimensional space corresponding to that which would have been found by the PCA. This linear MLP then performs a PCA which will be used as a basis for comparison with the behavior of a non-linear MLP. The only difference between the two networks is the non-linearity provided by the activation function. The aim is now to verify that this function detects a non-linearity in the data structure, and therefore performs a better cartography, which can be qualified as a non-linear projection of the clusters.

3.3. Determination of the activation function

The intrinsic data structure is unknown *a priori*. We are looking for a sigmoid function f which produces a better approximation of the data S_f at the output than the result obtained via the factorial plane (S_{id}). The comparison criterions that we define for evaluating the results are *MSE* and Quality. Quality is defined in terms of global relative error and is expressed as a percentage. Thus, for all m clusters, the global relative error (*GRE*) is:

$$GRE = \frac{\sum_{l=1}^{m} ISE(l)}{\sum_{l=1}^{m} \|X_l\|^2}$$

The quality is thus: $Qualiy = (1 - GRE) \times 100$. In the case of PCA, this corresponds to the quality of the data global projection on the first factorial plane, which is also called the percentage of explained inertia. Table 1 shows the results summarized in three columns: [1] Initial *MSE* = *MSE* before learning; for PCA, Initial *MSE* is comparable to the sum of the normalized squares of the vectors ; [2] Final *MSE* = *MSE* after learning; [3] Quality.

Table 1. Results obtained with different activation functions on the "prion diseases" data set

Function	Number of iterations by thousand	Initial *MSE*	Final *MSE*	Quality
PCA		0.8418	0.6922	17.77
x	100	0.8640	0.6962	17.29
f(x)	100	8.307	1.027	-22.00
$g_1(x)$	4000	0.8418	0.6943	17.52
$g_{15}(x)$	100	0.8842	0.6740	19.93
$g_{15}(x)$	160	0.8842	0.6694	20.47

The first row of table 1 presents the result obtained with PCA, and provides the basis for estimating the quality of the results obtained with the MLP. A study of table 1

allows us to verify that the identity activation function: $Id(x) = x$ gives results which are very close to the PCA. We then use the sigmoid activation function f. The function f brings out the problem that the output s is positive for all neurons, while the input x does not need to be (the data are centered). We introduce the additional term (-1/2) to make the function odd, and thus preserve positive and negative signs. By analogy with numerical methods used to find eigenvectors and eigenvalues, we introduce a factor *acc* to spread out the spectrum of the covariance matrix. The function thus defined is the function g_{acc} given by:

$$g_{acc}(x) = f(acc \times x) - \frac{1}{2} \text{ with } f(x) = \frac{1}{\left(1 + e^{-x}\right)}$$

The functions g_{acc} allows the MLP to obtain an initial *MSE* comparable to the initial *MSE* of the PCA (= 0.8418) while for the function f, the initial *MSE* is high (8.307). The function g_1 provides better results than the function *Id* but with a slow rate of convergence (4 million iterations). The factor *acc* speeds up the convergence: 160 thousand iterations with *acc=15*, moreover, this factor provides the lowest value of final *MSE* (0.6694), and the highest Quality of 20.47%. Thus we have adopted the function g_{15} which enables the optimal minimum obtained by the PCA to be passed, and to find a final *MSE* between a 2-dimensional PCA and a 3-dimensional PCA.
In conclusion, we can observe that the results obtained with the MLP are at least similar to the PCA results, therefore the learning process may be considered to be achieved. This process carries out a minimization whose result depends on the correlations between the data. Nervertheless, the large dimension of the data implies a low final quality with MLP as well as with PCA.

4. Mapping Knowledge Organization

Before examining the maps obtained by the two methods, we present an approach that we call Related Components Analysis, for validating and explaining the map obtained in a format better suited for human reading. Indeed, one of the objects of KDD is to generate a visualization of knowledge in a form suitable for verification or interpretation providing users with useful or interesting knowledge.

4.1. Related Components Analysis (RCA)

This method is based on graph theory. It defines the related components which represent the relative closeness between clusters. These related components are not defined according to predefined thresholds, but 10 proximity levels are calculated from the distances between clusters. The highest level is defined by the minimum distance between clusters and the lowest by the maximum distance between clusters. At a given level, two clusters are connected if their distance is lower than the maximum threshold of that level. Once the connections are calculated, sets of clusters linked up by a connection path, named "related components", are defined. This

operation is repeated for each level. While this method does not have the means to project the individual points (clusters), it clearly shows their closeness and separation in multidimensional space.

The maps obtained by PCA and MLP do not allow a complete representation of the position of the clusters, because their quality value is not superior to 21% as we can see in table 1. To compensate for this, we use the RCA. This technique gives the analyst the means of verifying if maps respect the distances between the clusters, and therefore the concentration of some clusters and the isolation of others. Moreover, the RCA facilitates the interpretation of the maps by allowing the clusters configuration to be visualized.

4.2. Presentation of the maps

The function g_{15} makes the MLP to converge on the plane giving the best account of the non-linearity. Once this plane has been determined, its axes can be interpreted in the same way as standard factorial axes. While this activation function gives a better approximation of the data than the PCA, it is less easy to visualize, because the clusters are close together at the centre of the space. To solve the visualization problem, we added a scale factor to the function g_{15} that is $g_{15}(scale*x)$. This operation is performed after the convergence of the MLP and can be thought of as unfolding the Kohonen map.

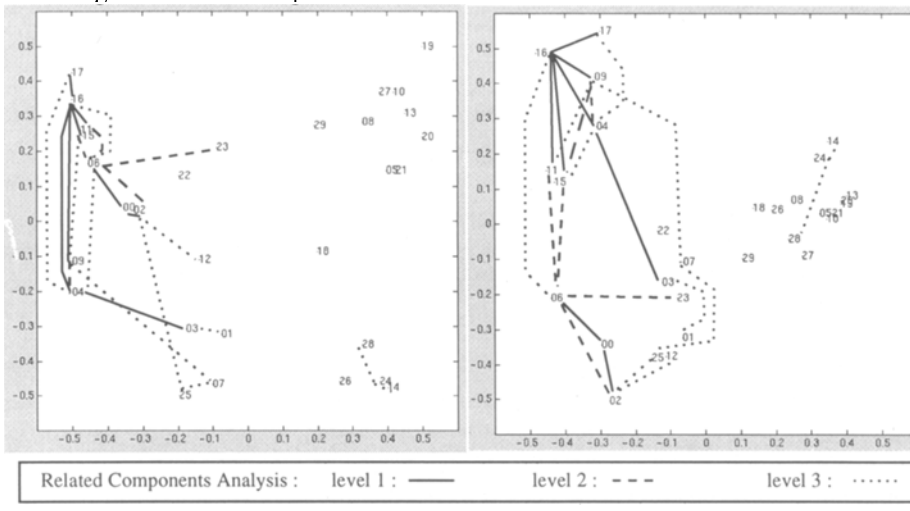


Fig. 2. MLP Map of the 30 clusters with function « g_{15} » and scale = 10

Fig.3. PCA Map of the 30 clusters. For the titles of clusters see table 1.

Figure 2 shows the map obtained by the MLP for our example of "prion diseases" with *scale=10*. Figure 3 shows the map obtained with the same set of data using the PCA. The related components of the three highest levels are shown on these figures. In figure 2, the left-hand side of the map groups together two very strong related components (level 1). According to our expert, these two components correspond to the domains of the "molecular biology of the prion protein" (component

[17,16,11,15,4,9]) and "scrapie" (component [6,0,2]). If we also look at the related components of levels 2 and 3, these two components are grouped together in a large set taking up the whole of the left-hand side of the map. This confirms the proximity of the two level 1 components. At the bottom right in figure 2, we note a level 3 related component (component [14,24,28]) which, according to our expert, corresponds to the problem of the "transmission of Creutzfeldt Jacob and Spongiform Encephalopathy Diseases". At the top right of the MLP map are the clusters more isolated in the multidimensional space. The map therefore respects the relationships between the clusters. In figure 3, the related components representing the "molecular biology of the prion protein" and the "scrapie" are also found on the left-hand side of the PCA map, but they are further away from each other, which is not justified by the RCA. The related component [14,24,28] representing "transmission of Creutzfeldt Jacob and Spongiform Encephalopathy Diseases", appears less well grouped on this map.
The above discussion shows that knowledge organization designate by the clusters can be represented in the forms of maps in which relations (proximities) can be examined by users or experts. A user-friendly hypertext interface allows to explore the knowledge organization interactively.

5. Conclusions and Open Problems

This study represent a step towards the development of a reliable automated mapping knowledge method. As mentioned in section 1, for representing the patterns of the inherent data structure, our approach applies cluster analysis and the geometrical representation of clusters in a 2-dimensional space in the form of maps. In this article, we have compared and evaluated two different cartography approaches: PCA and MLP. We have also explained in what manner a MLP, with the architecture and learning algorithm described in section 2, can be used to perform a PCA if the activation function is set to the identity function. This allowed us in section 3 to look for the non-linear activation function which best fits the data structure. The *MSE* was used as the means of comparison.
We have noticed that MLP detect a non linearity in the clusters structure that the PCA does not detect. In fact, the MLP has given a better performance than 2-dimensional PCA, but lower than 3-dimensional PCA. In other words, the MLP does not express the non linearity completely. A further study of the structure defined by the clusters would allow us to determine the degree of freedom of such structure, and to know whether it can be described by two factors. Such results could be used to create maps which better represent the knowledge content of the data.
The maps studied in this article are not only means of visualization. They also represent an analysis tool insofar as they allow us to evaluate the relative position of clusters (or topics) in the multidimensional space of representation. As observed in section 4, we must deal with the problems of readability of such maps. In the case of the MLP map, we have added a scale factor for improving its readability. A RCA, based on graph theory, was also used to show and evaluate best-fit distances and

proximities in multidimensional space, and compensating the approximate nature of the maps obtained by the MLP or the PCA.
Finally, as regards of the remark that KDD process involves "the evaluation and possibly interpretation of the patterns to make the decision of what constitutes useful or interesting knowledge and what does not" [4, p. 9], all that can be said currently is that generating maps easily understood by people is a promising field that calls for further studies. Furthermore, cartography may indicate an interesting new domain of research and development in knowledge discovery systems.

References

1. Bourret, P., Reggia, J., Samuelides, M.:Réseaux neuroneaux, une approche connexionniste de l'intelligence artificielle, Toulouse, TEKNEA (1991).
2. Brachman, R.J., Anand, T.:The Process of Knowledge Discovery in Databases, in Adavances in Knowledge Discovery and Data Mining. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. Editors. Menlo Park, Calif. AAAI Press / The MIT Press, (1996) 37-57.
3. Elder, J., Pregribon, D.: A Statistical Perspective on Knowledge Discovery in Databases, in Databases, in Adavances in Knowledge Discovery and Data Mining. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. Editors. Menlo Park, Calif. AAAI Press / The MIT Press (1996) 83-113.
4. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, in Adavances in Knowledge Discovery and Data Mining. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. Editors. Menlo Park, Calif. AAAI Press / The MIT Press (1996). 1-34.
5. Gallinari, P., Fogelman-Soulié, F.: Progressive Design of MLP Architecture", Neuro-Nïmes (1988) 171-182.
6. Gallinari, P., Thiria, S., Fogelman-Soulié, F.: Multilayer perceptrons and data analysis, International Conference on neural networks, IEEE 1 (1988) 391-399.
7. Grivel, L., François, C.: Une station de travail pour classer, cartographier et anlyser l'information bibliographique dans une perspective de veille scientifique et technique, in Noyer, J.M. Editor, Les sciences de l'information: Bibliométrie, Scientométrie, Infométrie, Rennes, SOLARIS II, Presses Universitaires de Rennes (1995) 81-112.
8. Holsheimer, M., Siebes, A.P. Data Mining: The Search for Knowledge in Databases. Amsterdam, CWI Report CS-R9406, ISSN 0169-118X. (1994)
9. Muller, Ch., Polanco, X., Royauté, J., Toussaint, Y.: Acquisition et structuration des connaissances en corpus : éléments méthodologiques. Rapport de recherche INRIA N° 3198, available in postcript format ftp.inria.fr (192.93.2.54) (1997)
10. Polanco, X., Grivel, L., Royauté, J.: How to Do Things with Terms in Informetrics: Terminological Variation and Stabilization as Science Watch Indicators, Proceedins of the Fifth International Conference of the International Society for Scientometrics and Informetrics. Edited by M.E.D. Koening and A. Bookstein. Medford, NJ: Learned Information Inc (1995) 435-444.
11. Polanco, X.: Extraction et modélisation des connaissances : une approche et ses technologies, Premières Journées du Chapitre Français de l'International Society for Knowledge Organization (ISKO), à Lille, France, les 16-17 octobre 1997 (1997)
12. Polanco, X., François, C., Keim, J-P.: Artificial Neural Network Technoloy for the Classification and Cartography of Scientific and Technical Information, Scientometrics, vol. 41, n° 1(1998) 69-82