Feature space transformation for transfer learning

Nistor Grozavu
LIPN-UMR 7030, University Paris 13,
99, av. J-B Clément, 93430
Villetaneuse, France
{nistor}@lipn.univ-paris13.fr

Younès Bennani
LIPN-UMR 7030, University Paris 13,
99, av. J-B Clément, 93430
Villetaneuse, France
{younes.bennani}@lipn.univ-paris13.fr

Lazhar Labiod
LIPADE, Paris Descartes University
45, rue des Saints Pères, 75006
Paris, France
{lazhar.labiod}@parisdescartes.fr

***Abstract*—In this paper, we propose a study on the use of weighted topological learning and matrix factorization methods to transform the representation space of a sparse dataset in order to increase the quality of learning, and adapt it to the case of transfer learning. The matrix factorization allows us to find latent variables, weighted topological learning is used to detect the most relevant among them. New data representation is based on their projections on the weighted topological model. Each object in the dataset is described by a new representation consisting of the distances of this object to all components of the topological model (prototypes).**

For transfer learning, we propose a new method where the representation of data is done in the same way as in the first phase, but using a pruned topological model. This pruning is performed after labeling the units of the topological model using the labels available for transfer.

The experiments are presented as a part of an International Challenge [1] where we have obtained promising results (5th rank).

I. Introduction

Data mining, or knowledge discovery in databases (KDD), an evolving area in information technology, has received much interest in recent studies. The aim of data mining is to extract knowledge from data. The data size can be measured in two dimensions, the size of features and the size of observations [2]. Both dimensions can be large, which may cause problems during the exploration and analysis of the dataset. Models and tools are therefore required to process data for an improved understanding [3]. Indeed, datasets with a large dimension (size of features) display small differences between the most similar and the least similar data. In such cases it is very difficult for a learning algorithm to detect similarity variables that define the clusters.

Features weighting is an extension of the selection process whereby features are assigned by continuous weights, which can be regarded as degrees of relevance. Continuous weighting provides more information about the relevance of various features. Clustering and features weighting are then clearly linked [4]. Applying these tasks in sequence can reduce the performance of the learning system. Therefore, a new algorithm for clustering and for features weighting is needed. Features weighting for unsupervised learning has received interest recently, and an interesting weighted method were proposed by Grozavu et al., called lwo-SOM [5] which represents an extension of the classical SOM algorithm [6] allowing to weight the relevant features. This approach allows to build a prototype matrix (to reduce the data size) and to weight the relevant features.

In this study, we focus on reducing the dimensions of the feature space as part of the unsupervised learning through the matrix factorization and the transformation of this space to facilitate the process of transfer learning.

The approximate factorization and tensor factorization (or decomposition) of a matrix have a main contribution in the improvement of data and the extraction of latent components. A common point for noisy detection, reduction of the model, the reconstruction of feasibility, and the BSS (Blind Source Separation) is to replace original data by an approximate representation of reduced dimensions obtained via a matrix factorization or decomposition.

The concept of matrix factorization is used in a wide range of important applications and each matrix factorization is a different assumption about the components (factors) of matrices and their underlying structures, and this choice is an essential process in each application domain.

Very often, the datasets to be analyzed are nonnegative (or partially positive), and sometimes they also have a sparse representation. For these datasets, it is better to take into account these constraints in the analysis and to extract the non-negative components or factors with physical meaning or a reasonable interpretation, and thus to avoid absurd or unpredictable results.

The singular value decomposition (SVD) treats the rows and columns in a symmetrical manner, and thus provides more information on the data matrix. This method also allows us to sort the information in the matrix so that, in general, the relevant part becomes visible. This property makes the SVD so useful in data mining and many other areas.

The bidiagonalisation GK (Golub-Kahan) method was originally formulated [7] for computing the SVD. This method can be also used to calculate a partial bidiagonalisation:
$AQ_k = P_{k+1}B_{k+1}$

where A is the data matrix, B_{k+1} are bidiagonal, and the clones Q_k and P_{k+1} are orthonormal.

With this decomposition, the approximations of singular values and singular vectors can be calculated similarly by tridiagonalisation. Indeed, it can be shown that the procedure of the GK bidiagonalisation is equivalent to applying the Lanczos tridiagonalisation on a symmetric matrix with a particular initial vector.

In our method we use this technique for the sparse data and Principal Component Analysis (PCA) for other datasets.

The rest of this paper is organized as follows. Section I presents briefly the principle of Matrix Factorization and the use of this technique for clustering, and the principles of the transfer learning. The methods proposed for the unsupervised learning and topological transfer learning are presented in Section II A and B. In Section III, we present the results of the validation and their interpretation. A conclusion and some perspectives are given in Section IV.

II. Transformation of the feature space

A. Unsupervised Transformation

The unsupervised learning is often used for clustering data and rarely as a data preprocessing method. However, there are many methods that produce new data representations from unlabeled data. These unsupervised methods are sometimes used as a preprocessing tool for supervised learning models.

Given a data matrix represented as vectors of variables (p observations and n features), the goal of the unsupervised transformation of feature space is to produce another data matrix of dimension (p, n') (the transformed representation of n' new latent variables) or a similarity matrix between the data of size (p, p). Applying a supervised method on the transformed matrix should provide better results compared to the original dataset.

The transformation of the feature space is done in two steps. First, we decompose the sparse data matrix using the SVD method. Then the matrix of latent variables obtained after this decomposition is used to learn a topological model (lwo-SOM [5]), to detect and weight the relevant features.

This approach uses the weights to filter the observation by adapting them during the learning process. Using this principle, the observation $\mathbf{x}$ is weighted (π) before computing the Euclidian distance and the objective function of lwo-SOM is presented as follows:

$$R_{lwo}(\chi, \mathcal{W}, \Pi) = \sum_{i=1}^{|E|} \sum_{j=1}^{|\mathcal{W}|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \|\pi_j \mathbf{x}_i - \mathbf{w}_j\|^2 \qquad (1)$$

where $\mathbf{x}_i$ represents an example (object) from the dataset, $\mathbf{w}_j$ is the prototype vector and π_j - the weights computed during the learning process. The final coding of each data point is based on the distances given for each of the prototype of the lwo-SOM model. This distance matrix represents the new description of the dataset. To assess the quality of this new data coding, the new representation is presented later in a classifier as a linear discriminant analysis (LDA).

For a training dataset A, an evaluation (test) dataset B, and a final evaluation (test) dataset C, the proposed method for feature space transformation is presented as following:

1) Normalization: $\widehat{A} = A * diag(std(A))^{\frac{1}{2}}$
2) Dimensionality reduction of the dataset $\widehat{A}$ by matrix factorisation: $svd(\widehat{A}) = [U_{\widehat{A}} S_{\widehat{A}} V_{\widehat{A}}]$

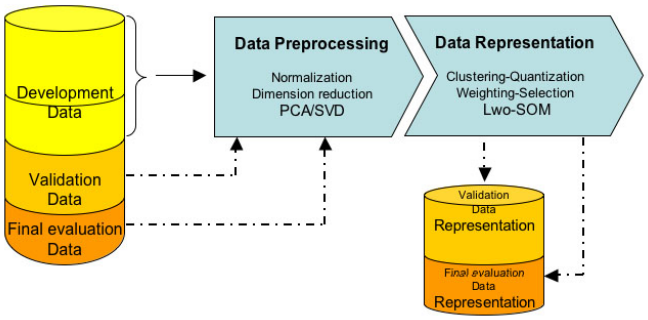


Fig. 1. Unspervised learning for feature space transformation

   For each column of $U_{\widehat{A}}$, $U_k = \frac{U_k}{\|U_k\|}$, where k is the number of retained eigenvectors
3) Matrix quantization: $P = lwo - SOM(U_{\widehat{A}})$
4) Apply steps 1 and 2 on the sets B and C
   $svd(\widehat{B}) = [U_{\widehat{B}} S_{\widehat{B}} V_{\widehat{B}}]$
   $svd(\widehat{C}) = [U_{\widehat{C}} S_{\widehat{C}} V_{\widehat{C}}]$
5) Calculation of distances matrices for B and C: $D = (d_{ij})$ where $d_{ij} = \|U_i - P_j\|^2$

Figure 1 illustrates the proposed process for unsupervised learning we used for the Challenge, and Figure 2 coresponds to the methodology used for the transfer of knowledge.

In the following (Algorithm 1) we present the proposed unsupervised learning algorithm for feature space transformation.

Algorithm 1: Transformation of the feature space and data coding

Inputs:
Learning (Training) data
Validation data
Final evaluation data
Output:
New representation of the validation and final datasets(the matrix decomposed)
Begin
Using the factorization of the initial matrix (training data)
Construct the prototypes matrix using the lwo-SOM algorithm
Construct the matrix of distances between the prototypes of lwo-SOM and validation and final evaluation matrices
End

B. Semi-supervised transformation

Predictive models capable to classify new objects (correctly predict the labels) generally require learning by using large amounts of labeled data.

Unfortunately, only a small amount of labeled learning data may be available because of the cost of manual annotation

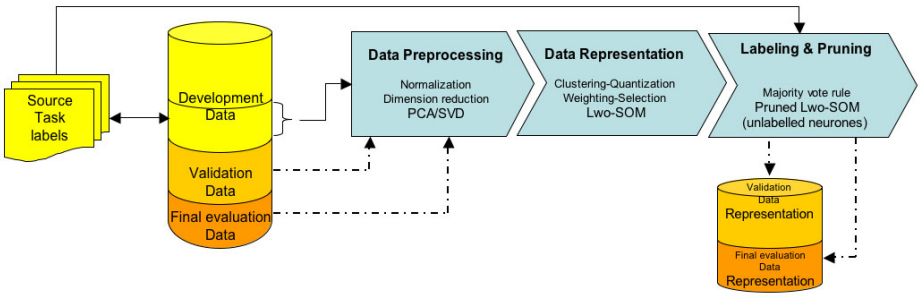


Fig. 2. Transformation of the feature space for Transfer Learning

of the data. Recent research has been focused on the use of large amounts of available unlabeled data, including: the transformation, the reduction of dimensionality, hierarchical representations of the variables (”deep learning”), kernel based learning, etc..

In some practical cases, it is desirable to produce representations of data that can be reused from one area to another. In this study, we examine how representation developed with a set of labels can be used to learn in an easier way another similar task. For example, in the field of handwriting recognition, labeled handwritten numbers are available for learning. The evaluation task would be the recognition of handwritten alphabet letters. We call this type of learning transfer learning.

For the transfer learning, we propose a new method for transforming the feature space where the representation of data is done in the same way as in the first unsupervised method, but using a pruned lwo-SOM map. This pruning is performed after labeling the prototypes matrix of lwo-SOM using available labels. Pruning is to remove all labeled prototypes (which represent labeled data) and we obtain a decomposition of the initial matrix, which resulted in the unlabeled prototypes matrix. Indeed, this new matrix, represents the data of other classes that are not available for transfer. These prototypes will be used as a dictionary for encoding validation data and final evaluation data. Indeed, the data validation and final evaluation sets are projected onto the unlabeled prototypes by calculating the Euclidean distance between these observations and prototypes of the lwo-SOM model. This distance matrix represent the new feature space and this new representation of the data is subsequently presented to a classifier such as the linear discriminant analysis (LDA).

The proposed algorithm for this second transformation (for the transfer learning) is presented in Algorithm 2.

Algorithm 2: Transformation of the feature space for transfer learning

Inputs:
Learning (Training) dataset
The knowledge to be transferred
Validation dataset
Final evaluation dataset
Output:
The pruned prototypes matrix
New representation of the validation and final evaluation datasets
Begin
Label the prototypes matrix (lwo-SOM map) using the available labels for transfer (majority voting rule)
Prune the map (the matrix of prototypes) by removing the labeled prototypes
Affect the validation data and final evaluation data on the final pruned matrix
end

III. Experimental protocol

Both proposed methods for the transformation of the data space have been tested as part of an International Challenge on Unsupervised and Transfer Learning [1]. The Challenge was made in two steps: The unsupervised learning for the transformation of the data space and the Transfer Learning. More details about the Challenge can be found on the official website of the Challenge (http://www.causality.inf.ethz.ch/unsupervised-learning.php). In the first phase of the challenge, no label is provided to participants. Participants are asked to produce representations of data that will be evaluated by the organizers in a supervised learning process (using labeled data that are not available for participants).

The transformed data should give better results in the supervised learning tasks used by the organizers to evaluate them. Labels for supervised learning tasks used for assessment remain unknown to the participants in Phase 1 and 2, but other labels will be available for transfer learning in Phase 2. In the second phase of the challenge (transfer learning), some labels are provided to participants for the same datasets used in the first phase, which will normally improve the representations of data obtained in the first phase.

Five datasets were available to participants in the Challenge (http://www.causality.inf.ethz.ch/unsupervised-learning.php).

Datasets:

- AVICENNA is a handwriting recognition dataset. The task of AVICENNA is to spot arabic words in an ancient manuscript to facilitate indexing. The data were formatted in a feature representation by the group of Mohamed Cheriet (Ecole de technologie superieure de Montreal, Quebec). The reception of this work is particularly intensive and widespread in the period between the late twelfth century to the first half of the fourteenth century, when more than a dozen comprehensive commentaries on this work were composed. These commentaries were one of the main ways of approaching, understanding and developing Avicennas philosophy and therefore any study of Post-Avicennian philosophy needs to pay specific attention to this commentary tradition.
- HARRY is a Human Action Recognition dataset. The HARRY dataset was constructed from the KTH human action recognition dataset and the Hollywood 2 dataset of human actions and scenes. The data include video clips shot on purpose to illustrate human actions (KTH data) and clips of hollywood movies (Hollywood2). The task is to recognize human actions like hand clapping, picking up a phone, walking, running, driving a car, etc. The data were preprocessed into a "bag" of STIP features.
- RITA is an image recognition dataset. This dataset was constructed from the CIFAR dataset that is part of the 80 million tiny image dataset. See this techreport, for details. The original data representation was enriched with new features and transformed to make the patterns unrecognizable.
- SYLVESTER is an ecology dataset. The task of SYLVESTER is to classify forest cover types. The forest cover type for 30×30 meter cells is obtained from US Forest Service (USFS) Region 2 Resource Information System (RIS) data.
- TERRY is a text recognition dataset. The data of TERRY come from a collection of Reuters, Ltd new articles made available by David D. Lewis: RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (12-Apr-2004 Version). The preprocessed data is a sparse representation based on a bag-of-word with a vocabulary of 47,236 stemmed tokens. Compared to the original dataset, the data were subsampled and scrambled and the features disguised.

Table 1 summarizes a description of datasets used to validate our approaches. All variables are numeric, and there are no missing values. Var. is the number of variables; Spars. - the percentage of sparse data, App. and transfer. are respectively the number of examples throughout the training data and the number of labels used to transfer during the transfer learning phase. The validation and final evaluation datasets consists of 4096 samples each.

The performance prediction are evaluated on the AUC curve and the area under the learning curve (ALC) on the test set versus the number of examples used to achieve learning (figure 3). Each curve consists of all the points for all used learning algorithms. The prediction performance was evaluated with the ALC (Area Under the Learning Curve).

TABLE I
DATASETS

Dataset	Domain	Var.	Spars.	App.	Transf.
AVICENNA	Handwriting	120	0%	150205	50000
HARRY	Video	5000	98.1%	69652	20000
RITA	Images	7200	1.1%	111808	24000
SYLVESTER	Ecology	100	0%	572820	100000
TERRY	Text	47236	99.8%	217034	40000

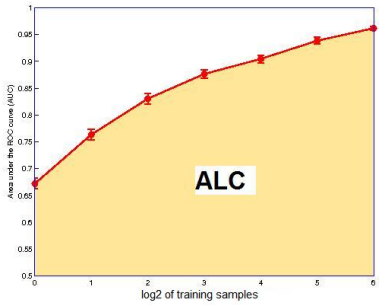
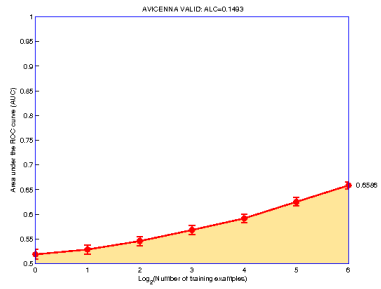
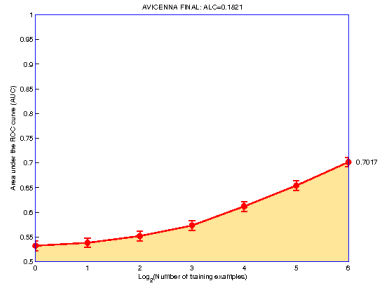


Fig. 3. ALC: Graphical representation



(a) Validation

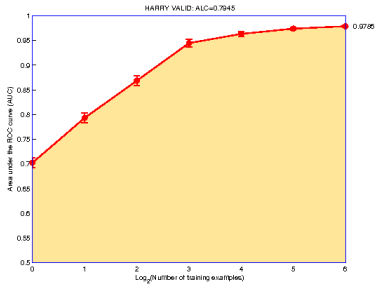


(b) Final Evaluation

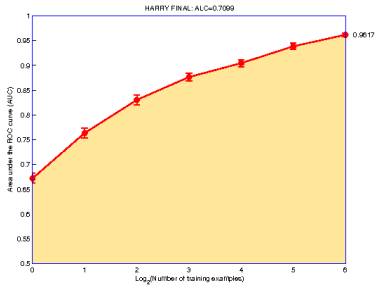
Fig. 4. AVICENNA: the AUC et ALC scores

The AUC (Area Under the ROC Curve [8] is calculated for all observations in the data set [9]. The obtained score is the standardized ALC calculated as follows:
$score = \frac{(ALC - Arand)}{(Amax - Arand)}$
where $Amax$ is the best achievable area (ie 1) and $Arand$ is

the area of a solution based on random predictions (or 0.5).

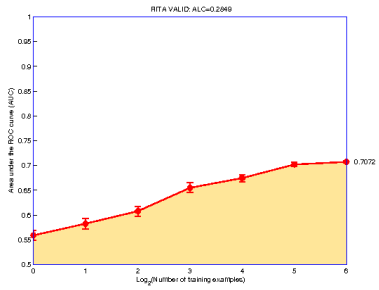


(a) Validation

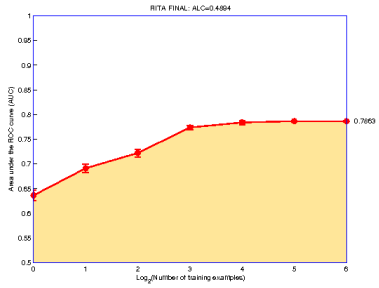


(b) Final Evaluation

Fig. 5. HARRY dataset: the AUC and ALC scores



(a) Validation



(b) Final Avaluation

Fig. 6. RITA dataset: AUC and ALC scores

For the AVICENNA dataset, we obtain a small AUC score which is 0.15 and 0.18 (Figure 4), but this is normal since AVICENNA is a difficult problem for the unsupervised learning. For this dataset our method retained 73 eigenvectors using PCA after the normalization, and a prototypes matrix of size 100 (10x10 cells). Almost all the participants of the challenge has the similar scores.

For the HARRY dataset, we built a prototype matrix size 900 (30x30 cells) by transforming the initial dataset using a matrix factorization technique (SVD) by retaining 20 eigenvectors. This allows us to obtain quite high AUC and ALC scores (Figure 5).

We used the SVD for the RITA dataset, and we built a prototype matrix size 900 (30x30 cells). The results made us ranked on the second place for this dataset in the Challenge (Figure 6).

After the dimensionality reduction of the SYLVESTER dataset using the PCA , we have built a matrix of prototypes of size 1600 (40x40 cells) and we got the AUC score of 0.61 for the validation dataset and 0.45 for the final evaluation dataset (figure 7).

Finally, for the TERRY dataset with 47236 features, we used the lwo-SOM method and we obtained a prototypes matrix of size 1089 (33x33 cells on the map) after a matrix transformation of the initial dataset using the SVD. The AUC and ALC scores and are given in Figure 8.

The results of the Challenge can be found on the oficial website Challenge (our team has the name NG-A3): http://www.causality.inf.ethz.ch/unsupervised-learning.php?page=results#cont

By analyzing the results of the Challenge, we can conclude that the approach we have proposed provides performance that exceed other methods such as: those based on Random Forests, factor analysis, the reduction of dimensionality with RBM, k-means type algorithm, etc.

Table II summarizes the AUC and ALC scores for the both Validation and Final Evaluation datasets using the proposed unsupervised learning algorithm for feature space transformation.

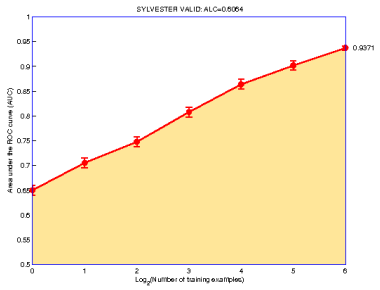
TABLE II
THE EXPERIMENTAL RESULTS FOR THE UNSUPERVISED LEARNING

Datasets	Validation		Final Evaluation	
	AUC	ALC	AUC	ALC
avicenna	0.658561	0.149326	0.701728	0.182106
harry	0.978586	0.794511	0.961722	0.709893
rita	0.707198	0.284878	0.786303	0.489439
sylvester	0.937103	0.606385	0.825077	0.44926
terry	0.990022	0.780955	0.994574	0.808953

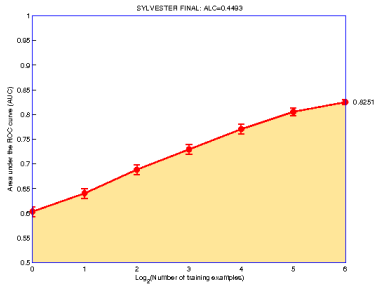
Contrarily, in the first phase of the Challenge, the winner (team name: AIO) [1] used a kernel based learning algorithm. Using the validation data, they have gradually improved the kernel. In Table IV, we show the results obtained by the AIO team and our team (NG-A3). As we can see, the score obtained using our proposed method is close to those obtained by AIO Team. Besides the Harry dataset, we get a higher ALC score.

All these results are summarized in Table III.

Table IV summarizes the results for transfer learning obtained using the proposed method.



(a) Validation



(b) Final Evaluation

Fig. 7. SYLVESTER: AUC and ALC scores

TABLE III
COMPARISON WITH THE BEST RESULT OF THE CHALLENGE

Method	Avicenna	Harry	Rita	Sylvester	Terry
AIO team	0.2183	0.7043	0.4951	0.4569	0.8465
Proposed Method	0.1821	0.7099	0.4894	0.4493	0.8089

IV. CONCLUSION

In this work, we proposed two methods for transforming the data features space: A method based on the combination of a matrix decomposition technique and a weighted topological learning, and an extension that uses a semi-supervised process for pruning the topological model. We adapted these methods for the Challenge "Unsupervised Learning and Transfer" to transform the feature space for different datasets. Our approaches have proven a high effectiveness for high dimensionality problems and different types of data.

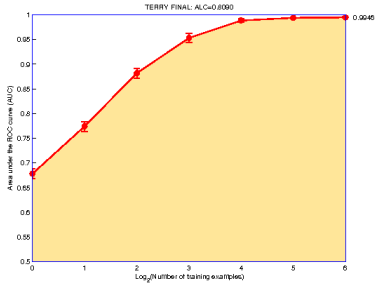
For the second phase of the Challenge, a learning methodology to transfer new knowledge has been proposed using a pruning technique of the matrix obtained with prototypes obtained with lwo-SOM.

The results are very promising and we were ranked 5th in the final ranking of the "Unspervised and Transfer Challenge"- NG-A3 team: http://www.causality.inf.ethz.ch/unsupervised-learning.php?page=results#cont.

TABLE IV
THE EXPERIMENTAL RESULTS FOR TRANSFER LEARNING

Datasets	Validation		Final Evaluation	
	AUC	ALC	AUC	ALC
Avicenna	0.637932	0.130236	0.623894	0.105119
Harry	0.978586	0.794511	0.961722	0.709893
Rita	0.707523	0.259007	0.759892	0.363303
Sylvester	0.936743	0.606771	0.624744	0.126217
Terry	0.983234	0.739909	0.888154	0.566029

(a) Validation



(b) Final Evaluation

Fig. 8. TERRY dataset: AUC and ALC scores

REFERENCES

[1] I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. W. Aha, "Unsupervised and transfer learning challenge," in *Proc. of International Joint Conference on Neural Networks 2011*, 2011.

[2] M. Verleysen, D. Franois, G. Simon, and V. Wertz, "On the effects of dimensionality on data analysis with neural networks," in *Artificial Neural Nets Problem solving methods*, ser. Lecture Notes in Computer Science 2687, J. A. e. J. Mira, Ed. Springer-Verlag, 2003, pp. II105–II112.

[3] V. Roth and T. Lange, "Feature selection in clustering problems," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2003.

[4] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(5), pp. 657–668, 2005.

[5] N. Grozavu, Y. Bennani, and M. Lebbah, "From variable weighting to cluster characterization in topographic unsupervised learning," in *in Proc. Proc. of IJCNN09, International Joint Conference on Neural Network*, 2009.

[6] T. Kohonen, *Self-organizing Maps*. Springer Berlin, 2001.

[7] G. H. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *in SIAM J. Numer. Anal.*, p. 205224, 1965.

[8] T. Fawcett, "Roc graphs : Notes and practical considerations for researchers," *Machine Learning*, 2004.

[9] C. Salperwyck and V. Lemaire, "Impact de la taille de lensemble dapprentissage : une tude empirique," *Confrence Internationale Francophone sur l'Extraction et la Gestion de Connaissance*, 2011.