Deep Self-Organising Maps for efficient heterogeneous biomedical signatures extraction

Nataliya Sokolovska, Nguyen Thanh Hai, Karine Clément, Jean-Daniel Zucker

Deep Self-Organising Maps for Efficient Heterogeneous Biomedical Signatures Extraction

Nataliya Sokolovska
UPMC University Paris 6
INSERM UMR_S 1166, NutriOmics Team
ICAN, Paris, France
Email: nataliya.sokolovska@upmc.fr

Nguyen Thanh Hai
UPMC University Paris 6
INSERM UMR_S 1166, NutriOmics Team
ICAN, Paris, France
Email: nthai@cit.ctu.edu.vn

Karine Clément
UPMC University Paris 6
INSERM UMR_S 1166, NutriOmics Team
ICAN, Paris, France
Email: karine.clement@psl.aphp.fr

Jean-Daniel Zucker
UPMC University Paris 6
ICAN, Paris, France
IRD, UMMISCO, Bondy, France
Email: jean-daniel.zucker@ird.fr

***Abstract*—Feature selection is used to preserve significant properties of data in a compact space. In particular, feature selection is needed in applications, where information comes from multiple heterogeneous high dimensional sources. Data integration, however, is a challenge in itself.**

In our contribution, we introduce a feature selection framework based on powerful visualisation capabilities of self-organising maps, where the deep structure can be learned in a supervised or unsupervised manner. For a supervised version of the deep SOM, we propose to carry out inference with a linear SVM. A forward-backward procedure helps to converge to an optimal feature set.

We show by experiments on real large-scale biomedical data set that the proposed methods embed data in a new compact meaningful representation, allow to visualise biomedical signatures, and also lead to a reasonable classification accuracy compared to the state-of-the-art methods.

I. Introduction

Heterogeneous data integration is a challenging task with an ambitious goal to increase performance of supervised learning, since various sources of data tend to contain different parts of information about the problem.

Structure learning and data integration allow to better understand the properties and content of biological data in general and of "omics" data (metabolomics, metagenomics, lipidomics, etc.) in particular. Combining complementary pieces issued from different data sources is likely to provide more knowledge, since distinct types of data provide distinct views of the molecular machinery of cells. Medical and biological knowledge can be naturally organised into hierarchies: symptoms of diseases are observed and pathological states on all levels of omics data are hidden. Hierarchical structures and data integration methods reveal dependencies that exist between cellular components and help to understand the biological network structure.

Graphical models follow a natural organisation and representation of data, and are a promising method of simultaneous heterogeneous data processing. Hidden variables in a graphical hierarchical model can efficiently agglomerate information of observed instances via dimensionality reduction, since fewer latent variables are able to summarise multiple features. However, integration of latent variables is a crucial step of modelling.

Multi-modal learning, heterogeneous data fusion, or data integration, involves relating information of different nature. In biological and medical applications, data coming from one source are already high-dimensional. Hence, data integration increases the dimensionality of a problem even more, and some feature selection or dimensionality reduction procedure is absolutely needed both to make the computations tractable and to obtain a model which is compact and easily interpretable.

Our goal is to develop an efficient feature selection approach which will design a compact model. The method needs to be scalable, to fusion heterogeneous data, and be able to reach a better generalising performance compared to a full model and to state-of-the-art methods. Another important question is whether introducing data of different nature have a positive effect, and provides additional knowledge. An important aspect of feature selection is whether a model is easily interpretable, and whether it is possible to visualise the results in order to investigate dependencies in the model.

In this paper, we propose a framework which is based on Self-Organising Maps (SOM) [15]. In this contribution, we run the learning procedure with linear support vector machines. The deep linear SVM has been considered and tested in [25], and it was reported that replacing the soft-max function in a deep structure by a linear SVM leads to a better accuracy. The learning procedure minimises a hinge or a margin-based loss, and not the cross-entropy.

Our contribution is multi-fold:

- we introduce and consider a simple deep feature selection framework which constructs layers of a deep structure layer-wise,

- the deep structure is based on the capability of SOM to visualise clustering in 2D,
- the proposed deep architecture can be learned either in a supervised or unsupervised mode,
- we illustrate that the proposed framework is efficient on a real original rich heterogeneous MicrObese data set [6], which contains meta-data, i.e., clinical parameters and alimentary patterns of patients, gene expressions of adipose tissue, and gene abundance of gut flora. We efficiently extract compact new data representations structured into a multi-level hierarchy. We evaluate the prediction power of the models with the reduced dimensionality showing that the proposed approach reaches the state-of-the-art performance.

The paper is organised as follows. Section II considers the related work and the state-of-the-art methods. We consider deep linear support vector machines in Section III and self-organising maps for feature selection in Section IV. We show the results of our experiments in Section V. Concluding remarks and perspectives close the paper.

II. Related Work

Learning features from unlabeled data is important in many applications [5], including bioinformatics and medical informatics, where the number of medical analysis or interventions is critical. Upper layers of hierarchical structures are more abstract data representations, whereas lower layers are low-level features from data. [1] states that optimisation in deep structures is not obvious. A possible explanation is that standard gradient-based approaches may easily get stuck near poor solutions. To learn a complex model efficiently, it is sometimes useful and beneficial to split a task into simpler models that can be estimated sequentially. The inference is extremely expensive in densely connected networks. Dimensionality reduction and feature selection is already a classical problem associated with deep structures (see [10] for an overview). Several heuristics have been proposed to make the problem tractable. Some of them are based on greedy layer-wise inference [9], [17], and the inference is reported to be quite efficient.

It has been shown [5] that feature selection with deep structures is sensitive to the number of hidden layers in graphs, and to the choice of an optimisation algorithm. So, in [5] it was demonstrated that a simple K-means clustering can provide very efficient new features representation (for an image processing task). When extracting features from plentiful unlabelled data, the dimension of a problem becomes easily very big. Apart from numerous feature selection methods, there are approaches how to deal with manifold. [22], e.g., proposed a classifier which is insensitive to local directions changes along the manifold.

The idea to do feature selection using SOM is not new. [12] introduced a simple greedy heuristic iterative approach for feature selection which includes 4 steps: 1) learn a SOM and label map; 2) if the classes overlap, then add a new feature or replace a feature; 3) if a feature does not improve the separability of the groups, eliminate or replace this feature; 4) re-train and re-label the map. We also propose a feature selection algorithm based on a clustering, and, namely, a SOM. Note, however, that [12] clusters observations and greedily looks for features ameliorating the separation of classes. We, on the contrary, cluster features, and look for best representatives in each feature cluster.

Clustering of features has been already considered by [3]. The principal interest was to build classifiers in a semi-supervised manner and to help analysts in their choice of features or attributes. Another motivation of [3] was to illuminate relationships between attributes, their relative importance for classification, and to better understand structure in data. Another clustering of features was done in [24]. [24] has introduced an algorithm called FAST which consists of two simple steps: 1) features are clustered (by using graph-theoretic clustering methods); 2) the most representative features somehow strongly related to classes are selected from each cluster to form a subset of new features. This approach is close to our idea. However, we do not estimate any relations to classes while choosing best representatives from clusters.

In this paper we use SOM clustering, however, it is possible to investigate the clustering with medoids for the same purpose. Partitioning around medoids (PAM) is introduced and described in details by [13], [14]. This is another quite efficient and robust clustering, which can be used for a hierarchy construction [23].

In an already classical deep architecture, in convolutional nets, the non-linearities are sigmoid functions, and new representations are learned in supervised mode using gradient descent. The advantages of the convolutional nets and SVM are combined in [11]. Deep structures learn complex mappings by transforming their inputs through multiple layers of nonlinear transformations [4]. There are several motivations to learn such an architecture. It is, first of all, a way to combine supervised and unsupervised learning. Second, there is a number of functions that can be used to compose weakly non-linear transformations. [4] introduced a multilayer kernel machines, which are based on an iterative procedure, repeated for each layer of a structure: 1) compute principal components in the feature space induced by a nonlinear kernel, and 2) prune components that are less informative from the feature space. Our approach, in its unsupervised mode, is a convolutional net. An interesting parallel between [4] and us, apart from using SVM, is that SOM is a nonlinear generalisation of the PCA.

Another avenue of research is controlling structure in data by penalty terms such as lasso-like methods. So, [2] proposed recently to add some convex constraints to the logistic regression penalised by the L_1 norm to produce a sparse model which involves a hierarchy restriction on feature interactions: an interaction is included if one or both features are marginally important. The disadvantage of the method is that the number of features in this approach is very big even for moderate-size applications, since the approach tests all interactions pairwise.

III. Deep Linear Support Vector Machines

To learn a hierarchical model, a training algorithm has access to n i.i.d. data points. We can either have labeled pairs $(X_i, Y_i)_{1\leq i\leq n}$, or an unlabelled data set $(X_i)_{1\leq i\leq n}$. The input variable or covariate is $X \in \mathcal{X}$, and the class variable is $Y \in \mathcal{Y}$, if the problem is supervised. The covariate variables are high-dimensional, and $X_i = (X_{i,1}, \ldots, X_{i,d})$, where d is the dimensionality of the problem. We are interested, in particular, to reduce the number of features in the model, so that the dimensionality of our problem becomes $r \ll d$, and so that we can carry out a classification task on a much more compact, and probably less noisy, feature space.

A deep structure can be learned with an SVM. A version of deep linear SVM which we exploit in our framework, has been introduced by [25]. The function in the linear case takes the following form:

$$\mathcal{L}(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N} \max(1 - \mathbf{w}^{\mathrm{T}}x_iy_i, 0), \tag{1}$$

and the rule to take a decision

$$\hat{y} = \arg\max_{y} \mathbf{w}^{\mathrm{T}}xy. \tag{2}$$

Let h_j be hidden or latent layers in the hierarchy, then

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial h_j} = -C\mathbf{w}y_j\mathbb{1}_{\{\mathbf{w}^{\mathrm{T}}h_jy_j>1\}}. \tag{3}$$

A way of producing probabilistic outputs from a kernel method (see [21]) is to use

$$p(y|x, \mathbf{w}) = \frac{\exp \mathbf{w}^{\mathrm{T}}x}{1 + \exp(\mathbf{w}^{\mathrm{T}}x)}. \tag{4}$$

The logistic function can be used as an activation function of a deep learning structure.

In such a deep supervised learning based on the linear SVM, we have two alternating steps. The forward step: apply logistic regression function to provide probabilistic interpretation and to activate units (weights are fixed). The backward step: compute the gradient presented as eq. (3), and update the weights. The gradient is one of the linear SVM, it is convex and differentiable, and we can apply any standard gradient descent method.

IV. Self-Organising Maps for Feature Selection

The Self-Organising Map (SOM) is an artificial network associated with the unsupervised learning paradigm [15]. It is famous for its efficient manner to map from a high-dimensional input space into a more compact space, usually to a two-dimensional output space. The two-dimensional representation is practical for a visualisation, since the mapping preserves topological relations between elements on the grid. Moreover, the continuous input space can be mapped into a discrete output space. The SOM belong to competitive learning methods, since neurons compete to be activated, and, as a result, only one is activated at a time. The winning neuron is called the winner. When the winner is set, all the other neurons have to re-organise themselves. Interestingly, the SOM can be seen as a non-linear generalisation of principal component analysis.

Given high-dimensional data $x \in \mathbb{R}^d$, the connection weights between observations i and the neurons of the grid j can be presented as $w_j = \{w_{ij} : j = 1, \ldots, K; i = 1, \ldots, n\}$, where K is the number of neurons on the grid. A discriminate function which is widely used, and which we also use in our experiments, is the squared Euclidean distance between an observation x and the weight vector w_j, for all j

$$D_j(x) = \sum_{i=1}^{n}(x_i - w_{ji})^2. \tag{5}$$

There are four principal steps of SOM learning shown as Algorithm 1.

Algorithm 1 Self-Organising Maps Learning Procedure

```
1. Initialisation: all connection weights are initialised randomly
2. Competition: for each observation, and all features, the neurons compute their values of a discriminant function. The neuron with the smallest value of the discriminate function is declared to be the winner
3. Cooperation: The winner determines the spatial location of a topological neighbourhood for other neurons, what provides the basis for cooperation between neighbouring neurons
4. Adaptation: Excited neurons decrease their values through an adjustment of the connection weights
```

The structure learning can be either supervised, or unsupervised.

A. Unsupervised Deep Self-Organising Maps

In an unsupervised setting, the feature selection procedure is completely unsupervised, and the algorithm performs only the first step, a forward pass. In this forward pass, we construct a deep structure layer-wise, where each layer consists of the clusters representatives from the previous level. A natural question which arises is whether such an unsupervised feature selection can be beneficial for a prediction task. Although it is currently impossible to provide a theoretical foundation for it, there is an intuition why a deep unsupervised feature selection is expected to perform and performs better in practice. Real data are always noisy, and a "good" clustering or dimensionality reduction can significantly reduce the noise. If features are tied into clusters of "high quality", then it is easier to detect a signal from data, and the generalising classification performance is higher. The hierarchical feature selection plays here a role of a filter, and a filter with multiple layers seems to perform better than a one-layer filter.

B. Supervised Deep Self-Organising Maps

The supervised deep SOM feature selection is based mostly on the forward-backward idea. Forward greedy feature selection algorithms are based on a greedily picking a feature at every step to significantly reduce a cost function. The idea is to progress aggressively at each iteration, and to get a model which is sparse. The major problem of this heuristic is that once a feature has been added, it cannot be removed, i.e. the forward pass can not correct mistakes done in earlier iterations. A solution to this problem would be a backward pass, which trains a full, not a sparse, model, and removes greedily features with the smallest impact on a cost function. The backward algorithm on its own is computationally quite expensive, since it starts with a full model [26].

We propose a hierarchical feature selection scheme with SOM which is drafted as Algorithm 2. The features in the backward step are drawn randomly.

Algorithm 2 Feature Selection with Forward-Backward SOM

```
// FORWARD
for each layer $l \in L$ // bottom up do
    Run a SOM
    Select representatives from each cluster to propagate them
    to an upper level
end for
// BACKWARD
for each layer $l \in L$ // top down do
    Estimate accuracy for level $l$
    Greedily update selected features
end for
```

V. Experiments

In this section, we describe our experiments and results on a real rich, and original biomedical data set. To construct the SOMs, we use *somtoolbox*[1] from Matlab. We also use SOM graphics from [18], [19], and [20].

A. Signatures of Metabolic Health

The biomedical problem of our interest is a real problem which is a binary classification of obese patients. The aim is to stratify patients in order to choose an efficient appropriate personalised medical treatment. The task is motivated by a recent French study [6] of gene-environment interactions carried out to understand the development of obesity. It was reported that the gut microbial gene richness can influence the outcome of a dietary intervention. A quantitative metagenomic analysis stratified patients into two groups: group with low gene gut flora count (LGC) and high gene gut flora count (HGC) group. The LGC individuals have a higher insulin-resistance and low-grade inflammation, and therefore the gene richness is strongly associated with obesity-driven diseases. The individuals from a low gene count group seemed to have an increased risk to develop obesity-related cardiometabolic risk compared to the patients from the high gene count group. It was shown [6] that a particular diet is able to increase the gene richness: an increase of genes was observed with the LGC patients after a 6-weeks energy-restricted diet. [16] conducted a similar study with Dutch individuals, and made a similar conclusion: there is a hope that a diet can be used to induce a permanent change of gut flora, and that treatment should be phenotype-specific. There is therefore a need to go deeper into these biomedical results and to identify candidate biomarkers associated with cardiometabolic disease (CMD) risk factors and with different stages of CMD evolution.

B. Brief Data Description

The MicrObese corpus contains meta-data, genes of adipose tissue, and gut flora metagenomic data. For each patient, we have the information to which class he or she belongs. There are two classes, high gene count (HGC) and low gene count (LGC) classes. Therefore, our problem is a binary prediction task from heterogeneous data.

In general, 49 patients have been hired and examined at the Pitié-Salpêtrière hospital, Paris, France [6], but as to the genes of the adipose tissue, we faced the problem of missing data, and not for all patients their class, LGC or HGC is provided. We decided to impute missing data by median values for the adipose tissue data. The patients who were not clearly stratified into the LGC or HGC group, were excluded from the analysis. Therefore, in our experiments we have access to 42 observations (patients). To get rid of important noise, after the discussion with pre-clinical researchers, we run a significance test (Kruskal-Wallis), and we keep those variables for which the raw (not adjusted for the multiple hypothesis testing) p-values < 0.05.

Figure 1 is a hierarchical structure based on SOM. Each upper layer is constructed from variables which are the closest ones to the unit centres of the pervious level. Here we also perform data integration. We carry out feature extraction for four data sources – metagenomic species, environmental data, host variables, and genes expressions for adipose tissue. We do feature selection separately for each data source (three layers). Then we integrate all selected variables in one analysis and obtain a mixed signature (also three layers). Taking into consideration that we would like to get a well-balanced signature, where each data type is presented by some features, the SOM of the lower levels of the hierarchy are constructed per data source, since the number of parameters are extremely different in, e.g., adipose tissue data and in the block of environmental variables. Although Figure 1 provides a schematic overview, the maps on the figure are exactly what we get in our experiments. It is interesting to see that lower levels where the number of parameters is quite big, do not reveal specific structures in data. The highest levels, on the contrary, show well-organised clusters. Figure 2 illustrates the quantisation error associated with hierarchies on each data sources and on the mixed hierarchy. It is easy to see that in all cases the quantisation error diminishes. Figure 3 illustrates the patients separation after the feature selection, where 1 stands for high gene count patients, and 2 for the low gene count ones. Note that each cluster may contain several patients.

The framework of Figure 1 can be applied to the whole MircroObese cohort, both to the HGC and to the LGC data points

[1] http://www.cis.hut.fi/projects/somtoolbox/

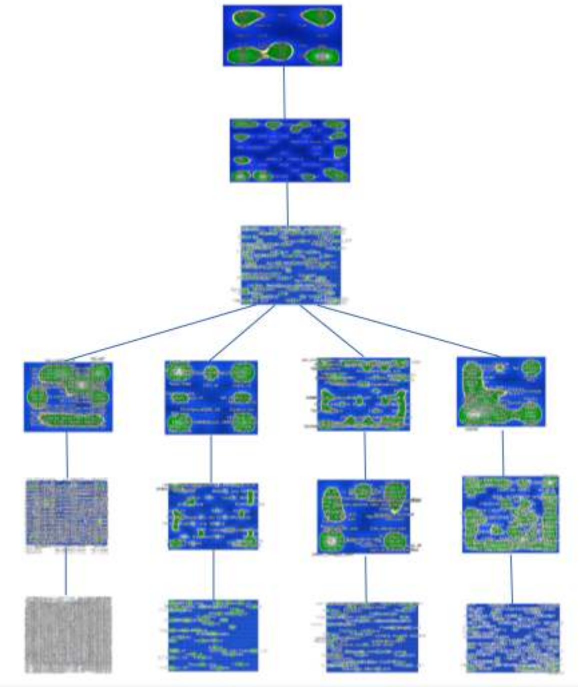

Fig. 1. The hierarchy of SOM. For three lower levels, from left to right: MGS, environmental variables, host, and adipose tissue microarray data. Three upper layers perform data integration from four data sources.

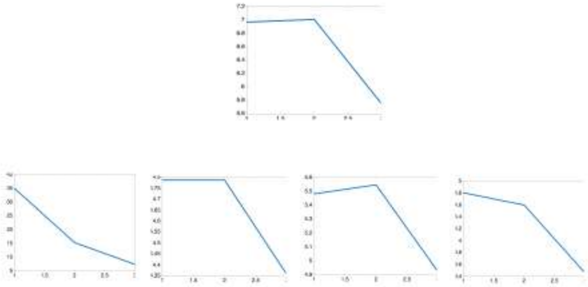

Fig. 2. Quantisation error. Above: the error associated with three highest levels of the SOM deep architecture; below: the quantisation error for each data source, associated with the lower three levels of the SOM hierarchy on Figure 1.

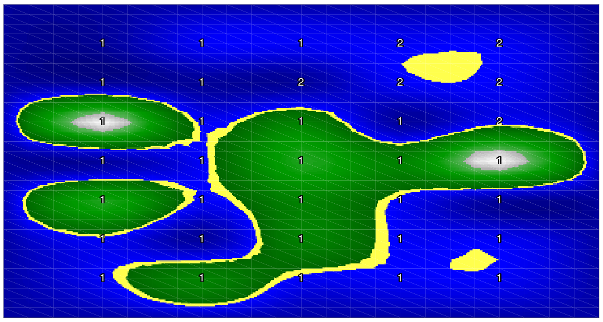

Fig. 3. Separation of patients with the selected features. 1– high gene count patients, 2 – low gene count patients.

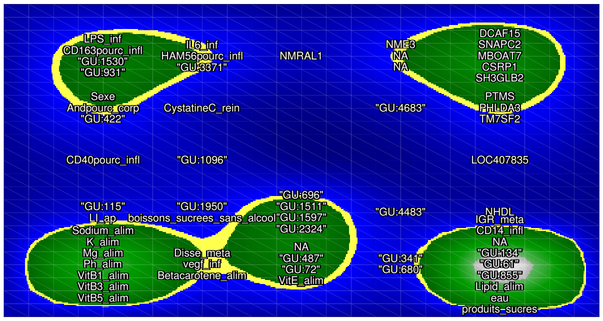

Fig. 4. Signature of the high gene count group which is associated with a better health.

(we do the 10-folds cross validation in all our classification experiments), but we can also split the data into the HGC and LGC data sets, and extract signatures for each group. These results that can be found on Figures 4 and 5 are very interesting for clinicians and researchers doing pre-clinical research, since these signatures allow them to better characterise the groups of patients.

Figure 5 shows the result of the prediction using the HGC and LGC groups. The signature, therefore, characterises the discrimination between two classes. It is a well-reported fact that biological and medical signatures are rather unstable. See, for instance, [8], where a comparison of more than thirty feature selection methods has been made, and where it has been shown that the stability of modern state-of-the-art approaches is rather low.

Another avenue to analyse signatures, is to construct Bayesian networks and to study the relations between the variables. We carry out feature selection with the deep SOM, and we run a Bayesian network on the selected variables. Figure 6 reveals the signature relations of the high gene count group, and Figure 7 of the low gene count group. The highest level of the deep SOM structure and the Bayesian networks provide complementary results. If we compare the relations for the HGC group, see Figures 3 and 6, we will see that the SOM clusters and the Bayesian networks quite often provide similar results, however, in some cases they reveal different relations between variables of interest. It is interesting, that the number of selected features for the LGC is bigger than for the HGC. Analysing Figures 4 and 6, we do the same conclusion: the SOM and the network reveal the same structure in data, with several interesting exceptions. The biomedical analysis of the results is out of scope of this paper, and will by done by fundamental biologists and clinicians.

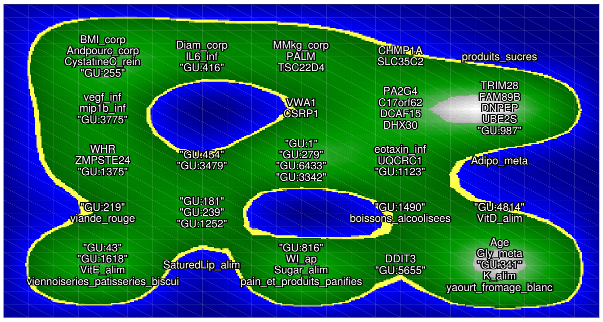

Fig. 5. Signature of the low gene count group which is associated with higher inflammation.

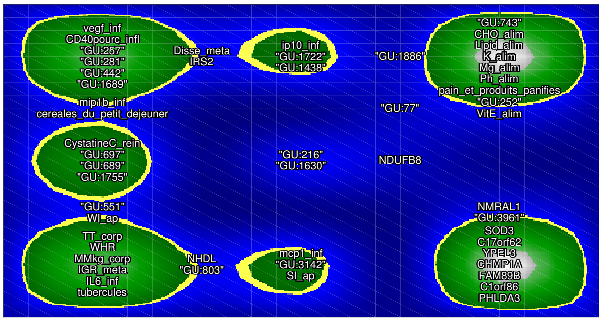

Fig. 6. A signature which discriminates high gene count and low gene count groups.

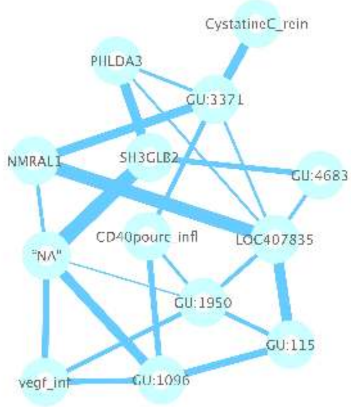

Fig. 7. Bayesian network of the selected features associated with the HGC.

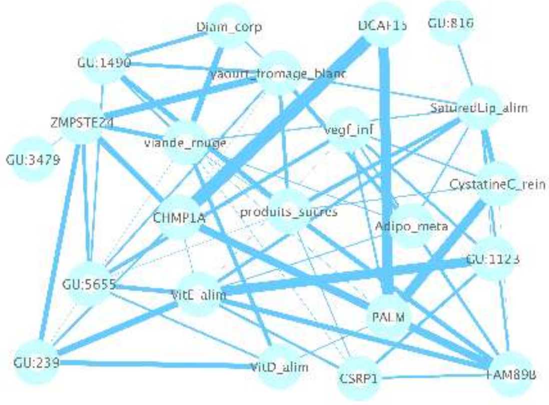

Fig. 8. Bayesian network constructed from the features selected for the LGC group.

C. Comparison with State-of-the-art Methods

In this section we show that the proposed feature selection method is efficient compared to the state-of-the art methods such as lasso and elastic net. In our experiments we use the *glmnet* R package [7]. Figures 9 and 10 show the 10-folds cross validation error rate on the MicroObese data as a function of the number of non-zero features in the model. We have done unsupervised feature pre-selection with the deep SOM, with the structure drafted on Figure 1. Then we apply the lasso to the pre-selected set of features, and compare the result with the lasso applied to the whole set of parameters (more than 2000 features).

Figure 9 on the left demonstrates the performance for the lasso applied to all features, without the unsupervised pre-selection step. On the right, we show the performance of our framework. Note that since the pre-selection step is unsupervised, we do not do any overfitting. The accuracy of both methods is comparable, and taking into consideration that the prediction task is quite challenging, the error rate around 0.3 – 0.33 is acceptable. However, the proposed framework applies lasso to a reduced data set, with a hundred of parameters, and not with thousands as the initial set.

Figure 10 illustrates our result which is similar to the one of Figure 9. It shows the performance as function of the number of selected features. On the left, we show the error rate for the elastic net ($\alpha = 0.5$ in the *glmnet* R package) on all features, and on the right, for the proposed approach. The elastic net achieves a higher accuracy than the lasso. Here (the plot on the right), as in the previous lasso experiment, we run the elastic net on about 100 variables, instead of thousands (the result on the left), and we see that the best model is one with about 30 parameters chosen among 100 from the features pre-selected in unsupervised manner. Note that the best model learned from all parameters (on the left), but with the comparable error rate, has more, namely, 43 features.

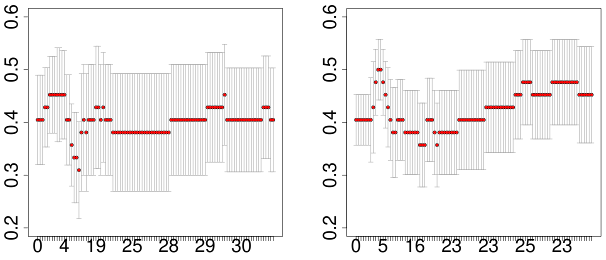

Fig. 9. The 10-folds cross validation error rate on the MicroObese data as a function of the number of active features. On the left: the lasso; on the right: the lasso after the unsupervised feature selection.

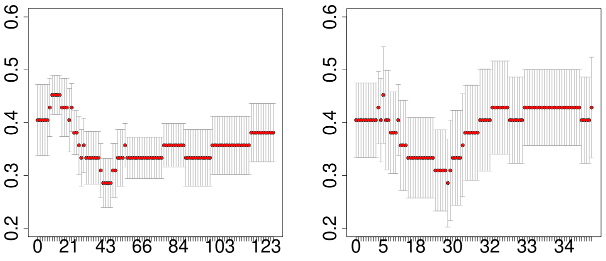

Fig. 10. The 10-folds cross validation error rate on the MicroObese data as a function of the number of active features. On the left: the elastic net; on the right: the elastic net applied to the data reduced by unsupervised feature selection.

VI. Conclusion

Data integration is a challenge, especially in applications where data are high-dimensional, e.g., metagenomics and the metagenomic species, and where the number of observations (patients) is small. We have proposed to reduce dimensionality by a deep approach which is based on self-organising maps, and which learns new compact data layer-wise, in a hierarchical way. We have considered supervised and unsupervised feature selection frameworks, as well as we considered a real data integration challenge. We show that the considered deep SOM approach is efficient on a real medical complex data set, and it is beneficial to combine it with the lasso and the elastic net approaches. The unsupervised feature selection diminishes the computational burden of the standard methods and also leads to the state-of-the art performance. Although the detailed biomedical discussion of the features clustering and of the quality of the obtained signatures is out of scope of this paper, and is to be done by biologists doing pre-clinical research, we expect that our framework can help to better stratify patients, and to develop methods of personalised medicine.

Acknowledgments

The clinical work was supported by Agence Nationale de la Recherche (ANR MICRO-Obes), KOT-Ceprodi and the association Fondation Coeur et Arteres. All ethical agreement are obtained. MicroObes Consortium list of contributors: Aurélie Cotillard, Sean P. Kennedy, Edi Prifti, Nicolas Pons, Emmanuelle Le Chatelier, Mathieu Almeida, Benoit Quinquis, Florence Levenez, Nathalie Galleron, Jean-Michel Batto, Joel Doré, Stanislav Dusko Ehrlich, Catherine Juste, Tomas de Wouters. This work is also part of the European Union's Seventh Framework Program under grant agreement HEALTH-F4-2012-305312 (MetaCardis project).

References

[1] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007.

[2] J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *The annals of statistics*, 41(3):1111–1141, 2013.

[3] R. Butterworth, G. Piatetsky-Shapiro, and D. A. Simovici. On feature selection through clustering. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM, 2005.

[4] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *NIPS*, 2009.

[5] A. Coates, A.Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research – Proceedings Track*, 15:215–223, 2011.

[6] A. Cotillard and al. Dietary intervention impact on gut microbial gene richness. *Nature*, 500:585–588, 2013.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2008.

[8] A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 12(6), 2011.

[9] G.E. Hinton, S. Osindero, and Y.W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[10] G.E. Hinton and R. Salakhudinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[11] F. J. Huang and Y. LeCun. Large-scale learning with svm and convolutional for generic object categorization. In *CVPR*, 2006.

[12] J. Iivarinen, K. Valkealahti, A. Visa, and O. Simula. Feature selection with self-organising maps. In *ICANN*, 1994.

[13] L. Kaufman and P. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, 1987.

[14] L. Kaufman and P. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley, 2009.

[15] T. Kohonen. The self-organising map. In *Proceedings of the IEEE*, 1990.

[16] E. Le Chatelier and al. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 2011.

[17] H. Lee, R. Gross, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

[18] E. Pampalk. Islands of music – analysis, organization, and visualization of music archives. *Journal of the Austrian Society for Artificial Intelligence*, 22(4):20–23, 2003.

[19] E. Pampalk, W. Goebl, and G. Widmer. Visualizing changes in the structure of data for exploratory feature selection. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[20] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *ACM Multimedia*, 2002.

[21] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[22] S. Rifai, Y. Dauphin, V. Pascal, Y. Bengio, and X. Muller. The manifold tangent classifier. In *NIPS*, 2011.

[23] N. Sokolovska, S. Rizkalla, K. Clément, and J.-D. Zucker. Continuous and discrete deep classifiers for data integration. In *International Symposium on Intelligent Data Analysis*, 2015.

[24] Q. Song, J.Ni, and G.Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):1–14, 2005.

[25] Y. Tang. Deep learning using linear support vector machines. In *ICML 2013: Challenges in representation learning workshop*, 2013.

[26] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *NIPS*, 2009.