ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Making machine learning models interpretable

Alfredo Vellido[1], José D. Martín-Guerrero[2]
and Paulo J.G. Lisboa[3] *

1- Dept. de Llenguatges i Sistemes Informàtics - Univ. Politècnica de Catalunya
C. Jordi Girona, 1-3, 08034 Barcelona - Spain

2- Dept. d'Enginyeria Electrònica - Universitat de València
Av. de la Universitat, sn, 46100 Burjassot (València) - Spain

3- Dept. of Mathematics and Statistics - Liverpool John Moores University
Byrom St. L3 3AF Liverpool - United Kingdom

Abstract. Data of different levels of complexity and of ever growing diversity of characteristics are the raw materials that machine learning practitioners try to model using their wide palette of methods and tools. The obtained models are meant to be a synthetic representation of the available, observed data that captures some of their intrinsic regularities or patterns. Therefore, the use of machine learning techniques for data analysis can be understood as a problem of pattern recognition or, more informally, of knowledge discovery and data mining. There exists a gap, though, between data modeling and knowledge extraction. Models, depending on the machine learning techniques employed, can be described in diverse ways but, in order to consider that some knowledge has been achieved from their description, we must take into account the human cognitive factor that any knowledge extraction process entails. These models as such can be rendered powerless unless they can be *interpreted*, and the process of human interpretation follows rules that go well beyond technical prowess. For this reason, interpretability is a paramount quality that machine learning methods should aim to achieve if they are to be applied in practice. This paper is a brief introduction to the special session on *interpretable models in machine learning*, organized as part of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. It includes a discussion on the several works accepted for the session, with an overview of the context of wider research on interpretability of machine learning models.

1 Introduction

The advent of the digital age has made almost any human endeavor the source of ever-increasing amounts of information. This information often takes the form of computable data, that is, data which are available in a format that can be machine processed and, ultimately, reasoned upon.

This data deluge pervades most scientific areas. A clear example of this can be found in bioinformatics and biomedicine. Even if the human genome was decoded only about a decade ago, genomic science has since become an almost

*This research was partially funded by Spanish MICINN TIN2009-13895-C02-01 project grant.

fully data-driven area. This can also be said about many other areas in biology research. In all of these, an army of new data-acquisition technologies coalesce with a widening range of investigation scales, from the molecule to the population, to make them a major challenge for intelligent data analysis [1]. The blossoming *-omics* sciences (genomics, proteomics, metabolomics and the like), in particular, have become a main target for machine learning researchers precisely because their dependency on large and non-trivial databases. As explicitly stated in [2] "[...] *the need to process terabytes of information has become de rigueur for many labs engaged in genomic research*".

The example of medicine is not too dissimilar. The commodification of healthcare, both in the public and private health sectors, is leading to a rapidly increasing demand for personalization of patients' treatments, which requires a sophisticated management of information systems [3]. This is, amongst other reasons, because the amount of medical information available to experts is increasing exponentially.

If interpretability is a necessary requirement in medical applications, it is no less relevant, for instance, in business applications and processes. Large operational databases are commonplace in retail and industry, and machine learning techniques are expected to transform these data into meaningful business knowledge in the form of actionable business plans. As bluntly stated in [4], a business manager "*is more likely to accept the [machine learning method] recommendations if the results are explained in business terms*". This comes to explain, for instance, the success of rule induction methods in this application field.

All in all, data of different levels of complexity and of ever growing diversity of characteristics are the raw materials that machine learning practitioners model using the battery of methods at their disposal. The obtained models are meant to be a formal representation of the available, observed data, for instance described as some formalization of the relationships between the data features. In one way or another, a model is meant to capture some of the intrinsic regularities or patterns that might be present in the data. Therefore, the use of machine learning techniques for data analysis can be understood as a problem of pattern recognition or, more informally, of knowledge discovery and data mining.

There exist a gap, though, between data modeling and knowledge extraction that should not be ignored. Models, depending on the machine learning techniques employed, can be described in diverse ways but, in order to consider that some knowledge has been extracted from the raw data, the human cognitive factor that any knowledge extraction process entails must be taken into account. Deductive reasoning is at the core of the scientific method, but inductive reasoning can be equally fruitful [5]. Nonetheless, in order to enable inductive reasoning from the results obtained by machine learning and related methods, humans need to resort to verbal and visual metaphors and the use of these metaphors opens the door to subjectivity, which is not always a desired scenario. Although interpretation and subjectivity cannot be extricated, it has been shown that the weight of preconceptions and prior beliefs in data and model interpretation can be at least partially assessed and controlled [6].

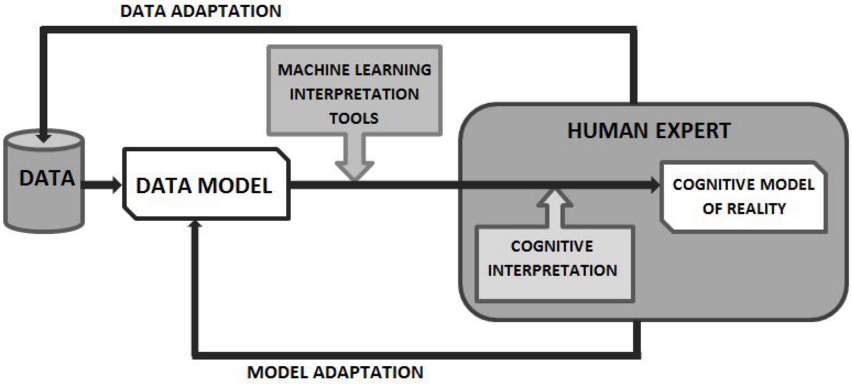

Fig. 1: An schematic graphical representation of the process of interpretation for machine learning models.

In any case, machine learning-based data models, regardless how sophisticated, can effectively be rendered powerless unless they can be *interpreted* by human experts, and the process of human interpretation does not necessarily match that of machine learning algorithms, since it follows rules that go well beyond technical prowess. For this reason, this tutorial poses the proposition that interpretability is a paramount quality that machine learning methods should aim to achieve if they are to be applied in practice.

The achievement of interpretability in data analysis using machine learning methods can be seen as a process with interacting stages, as described schematically in Figure 1. Data models are generated using machine learning tools and these are interpreted using methods that are tailored to these tools. Then, the human interpretation of these results must be communicated in the language of the domain expert and can feed back onto the process by advising either data or model adaptation.

Machine learning interpretability is thus the theme of a special session at the 2012 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, for which this brief paper is a tutorial. In section 2, we focus on the theme of dimensionality reduction as an efficient approach to model interpretation in machine learning, while, in section 3, we provide an overview and discussion of the different machine learning interpretation approaches proposed by the authors of the papers presented in this session.

2 Interpretation through dimensionality reduction

The size of currently available databases makes scalability a necessary requirement for real world applications. Sophisticated adjustments to existing machine learning techniques are often required to achieve this goal [7].

It is not uncommon to find that this problem concerns not only the number of cases available in the database, but also the number of data attributes. Problems of high and very high dimensionality (that is, problems in which the analyzed datasets consist of hundreds or even thousands of variables) are becoming commonplace in industry and bioinformatics, amongst other areas (think, for instance, of the analysis of microarray data in genomics, where thousands of variables, many of them likely to be uninformative, must be considered [8]).

Almost no problem is interpretable in practice if all data attributes are retained and used to provide an outcome. Beyond practicality, when such a large number of attributes is available, many of them are likely to be irrelevant to the outcome of the method, if not directly counterproductive. Furthermore, data of very high-dimensionality are bound to show unexpected geometrical properties that might bias the interpretation of results [9].

The presence of a large number of attributes is often tackled using methods of dimensionality reduction (DR). There are two main DR approaches available to the analyst: feature selection (FS), in which features are appraised individually in order to either retain or discard them [10], both for supervised and unsupervised [11] problems; and feature extraction (FE) [12], in which new non-observable features are created on the basis of the original, observed ones.

It is worth noting that some of the most popular DR techniques in real-world applications are precisely some of the simplest ones. This comes to explain, for instance, the resilience and widespread use of a more-than-a-century old linear FE technique such as Principal Component Analysis (PCA, [13]). It is not just simple, but also readily interpretable, because the extracted features are linear combinations of the observed ones and, as a result, the outcome can still be intuitively explained in terms of the latter. Moreover, it allows a very straightforward data visualization through data projection onto the main extracted components.

An example of the power of a simple FS method can be found in [14], where a basic but thorough backward selection procedure on top of a linear Single Layer Perceptron (SLP) model achieved both high accuracy and maximum interpretability in a medical problem of classification of human brain tumours, using only three variables (selected out of almost two hundred) to discriminate between two types of malignant tumours. In this problem, as in many more involving the application of machine learning methods in medicine [15], the use of FS is almost compulsory. Medical experts will only accept a parsimonious outcome from a machine learning method, as they require a clearly explainable basis for their decision making tasks that, furthermore, complies with their standard operational guidelines, often based on simple and rigid attribute scores. The alternative of FE often remains out of bounds, even if suitable to the problem at hand, unless it can be easily reverted to the original observed variables.

With this in mind, it is also the case that some of the most interesting and relevant machine learning contributions to the problem of multivariate data DR have stemmed from the field of nonlinear dimensionality reduction (NLDR) [9]. The challenge of interpretability is very explicit here: nonlinear methods rarely provide an easy interpretation of the outcome in terms of the original data features, because such outcome is usually a non-trivial nonlinear function of these features.

NLDR techniques usually attempt to minimize the unavoidable distortion they introduce in the mapping of the high-dimensional data from the observed space onto lower-dimensional spaces. Many approaches to this problem have been presented, but they are beyond the scope of this paper. Some are reviewed elsewhere [16] and the own ESANN conference has devoted special sessions to this problem [17]. The problem of finding the adequate output dimension for NLDR techniques is an area of research on its own [18]. Likewise, much effort has been made to embed the NLDR projection or mapping distortion into the own machine learning training process, for instance in the form of *magnification control* [19, 20].

The pursuit of interpretability in NLDR methods is still a wide open and most interesting research challenge. Dimensionality reduction, in its most extreme form, can lead to methods of information visualization. As stated in the introduction, interpretability may be seen as a problem of knowledge extraction from data regularity patterns. One of the forms in which we can achieve knowledge extraction is precisely through visualization. As stated in [21], information visualization can help us to gain insights into a problem through graphical metaphors, in a uniquely inductive manner that taps into the sophisticated visual capabilities of human cognition. The visualization of multivariate data involves a problem that goes beyond artificial pattern recognition using machine learning and related techniques to involve a proactive observer. The cognitive processing of visual stimuli [22, 23] includes an element of natural subjectivity, that the analyst must aim to control as much as possible [6].

Many of the most relevant recent machine learning contributions to multivariate data visualization have their origin in the field of NLDR [9]. A well-know and widely used NLDR method for data visualization in low-dimensional spaces is Kohonen's Self-Organizing Map (SOM) [24], in its many variants. This method attempts to model data through a discrete version of a low-dimensional manifold consisting of a topologically ordered grid of cluster centroids.

The nonlinearity of a method such as SOM entails the existence of local distortion (magnification) in the mapping of the data from the observed space onto the visualization space. This involves nonlinear manifold stretching and compression effects that limit the direct interpretation of the visual data representation. Its nonlinearity has not prevented SOM to achieve mainstream status, even in very practical application fields [25]. In any case, the nonlinear distortion introduced by an NLDR method such as this is still problematic from the viewpoint of the achievement of interpretability. There have been efforts to provide visual solutions to this limitation [26], by defining and visualizing DR quality measures that, embedded in the method, can be associated to each data point, using coloring procedures for the data-corresponding cells in the Voronoi tesselation [27] of the projection space.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

3 Making Machine Learning interpretable: Contributions to the 20th ESANN special session

A total of nine papers were accepted for the special session on *interpretable models in machine learning*, organized as part of the 20th ESANN conference. All these papers provided diverse and insightful methods to address the problem of interpretability for a number of different machine learning techniques. They also differed in scope: some of them are purely theoretical, whereas others are very specifically application-oriented. Several focused on interpretation through visualization methods, and that is the reason why we paid special attention to the problem of DR for visualization in the previous section. What follows is a brief structured discussion of their contributions.

In traditional statistics, an attractive way of presenting models to non-expert mathematicians is via graphical tools. An arguably old-fashioned but good example of such a tool is the nomogram. When talking about non-parametric models, the weights of the different covariates are not constant, which is a barrier for the use of this graphical approach in machine learning. Nevertheless, some progress along this direction is shown in [28], in the context of the medical problem of survival modeling. This work sets for itself the task of making some *black-box* models (support vector machines) interpretable, which is accomplished using constant B-spline kernel functions and sparsity constraints. The challenge of interpretability for nonlinear machine learning methods, in the context of the study, is clearly stated by the authors: "*clinicians are interested in decision support supplied without interfering with the clinical work flow, in an automatic way and providing recommendations*".

However, and as explained in the previous section, nonlinear models can have locally linear properties. These can be used to focus on the most relevant degrees of freedom captured by the model. This leads to a wide range of approaches to NLDR, as described in the previous section. One sensitive aspect of NLDR methods is that of defining adequate evaluation measures to assess their performance. Many of these come under the umbrella of the co-ranking framework described in [29]. A new improved parametrization for this framework is presented in this session in [30]. Importantly, this is linked to easy-to-visualize point-specific quality measures. The advantage of using locally-linear, globally-nonlinear models is shown in [31], where Fuzzy-Supervised SOM (FSSOM), a semi-supervised variant of SOM that takes into account class label information, is applied to a problem of hyperspectral images unmixing.

Some NLDR methods, such as SOM and Generative Topographic Mapping [32], allow an explicit quantification of the local distortion of the mappings they generate in order to achieve low-dimensional visual data representations. In [33], a cartogram-based [34] method to reintroduce the local distortion into the low-dimensional data visualization provided by the batch-SOM algorithm is provided. By reintroducing this distortion explicitly, the non-linearity of the mapping is factored in the visualization, which should help to ease its interpretation.

An alternative to visualizing the geometry of the data distribution is by recourse to mapping analytical classifiers into sets of explanatory rules that apply to distinct sub-cohorts of data [35]. This approach has the advantage of speaking the language of experts and so can be an important step in validation tests in hazard and operability analysis (HAZOP), that is to say, to verify that the classifier is "doing the right thing" by using the right variables in the right way, by checking against prior knowledge. A further interpretation is in helping to diagnose the model performance, where unexpected correct or incorrect classifications may be attributed, for instance, to outliers and data artifacts.

In [36], we find a basic instance of this approach, in which visualization aids are provided for classification trees, a type of methods favored in application fields such as business [4]. These aids target the input data distribution for each class in each terminal node, using a method called Sectors-on-Sectors that builds on work presented by the same authors at ESANN 2011 [37].

Increasingly, there is a tendency to marry rule-based interpretation with direct analytical inference of the posterior probability of class membership. This may be done using reference cases, the way a clinician, for instance, may interpret new cases by reference to particular prototypical examples. The analytical approach to model interpretation is by generating such prototypes.

The central concept in this approach is that of similarity or dissimilarity between individual data points. These measures need to be set in the context of the posterior probability distribution. Several of the papers in the current special session relate to this [38, 39, 41]. In [39], expert medical knowledge is integrated in a Fuzzy Supervised Neural Gas model (similar to the one used in [31]) by explicitly coding such information into a class similarity/dissimilarity measure, then used in classification to judge class label agreement. The practical interpretability of the resulting model increases through the integration of this expert-generated information.

The integration of knowledge (in this case in the form of biological information from genomic databases) in a machine learning process as a way to increase model interpretability is also followed in [40]. The authors discuss Structured Variable Selection (SVS), a machine learning-based *pipeline* for the analysis of high-throughput data that includes a step of semantic clustering and visualization. This step increases the interpretability of the results through the identification of their biological meaning.

A particular approach is to embed the statistical geometry of the posterior distribution into the data space by calculating a nonlinear metric from which similarity between data points with respect to classification probability is reflected in the geodesic distance between them. In this special session, such approach is followed in [41]. The disadvantage of this approach is that the metric is non-Euclidean, therefore projective methods often used in data visualization do apply directly. This enables the whole data set to be represented in a single network, from which communities or other structural properties of the data may be inferred. Moreover, the pairwise geodesic distances can be mapped onto a rigorous Euclidean space where standard projective approaches to data visualization will also apply. An example of this is the application of semi-supervised blind signal separation, for instance with convex-NMF [42], which is reported elsewhere [43].

An orthogonal direction to visualizing the space of observations is to visualize instead the relationship between the covariates. This leads to multivariate association maps, also known as graphical models, which can be useful for gaining insights into the data structure. An extension of this approach is the derivation of causal models, as described in ESANN 2011 in [44].

A different problem in machine learning that also involves interpretability issues is that of defining interpretable machine learning methods capable of dealing in a consistent way with data of heterogeneous nature. A two-layer artificial neural network is presented in [38], in which the neuron model computes a similarity function between data inputs and model weights. This model is capable of coherently analyzing variables of different nature: continuous, ordinal, and categorical, even if information is partially missing.

All of the above are generic approaches that may be applied to static data. For time series, and more specifically, for failure time data, suited to longitudinal data analysis, specific statistical considerations apply due to the occurrence of censorship. In addition, the insights generate by these models are inherently set in the time domain. Going back to the work in [28], the authors' approach provides an example of how this can be done, complementing the approach of directly modeling the failure rate, i.e. the hazard distribution [45].

References

[1] P.J.G. Lisboa, A. Vellido, R. Tagliaferri, F. Napolitano, M. Ceccarelli, J.D. Martín-Guerrero, E. Biganzoli, Data mining in cancer research, *IEEE Computational Intelligence Magazine*, 5(1):14-18, 2010.

[2] S.D. Kahn, On the future of genomic data, *Science*, 331(6018):728-729, 2011.

[3] A. Vellido, E. Biganzoli, P.J.G. Lisboa, Machine learning in cancer research: implications for personalised medicine. In M. Verleysen, editor, *proceedings of the 16th European Symposium on Artificial Neural Networks* (ESANN 2008), d-side pub., pages 55-64, Bruges (Belgium), 2008.

[4] I. Bose, R.K. Mahapatra, Business Data Mining - a machine learning perspective. *Information & Management*, 39:211-225, 2001.

[5] D.B. Kell, S.G. Oliver, Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26(1):99-105, 2004.

[6] C. Ziemkiewicz, R. Kosara R, Preconceptions and individual differences in understanding visual metaphors, *Compututer Graphical Forum*, 28(3):911-918, 2009.

[7] R. Collobert, *Large Scale Machine Learning.* Ph.D. Thesis, Université de Paris VI, Paris, France, IDIAP-RR 04-42, 2004.

[8] P. Baldi, S. Brunak, *Bioinformatics: The Machine Learning Approach.* MIT Press, 2001.

[9] J.A. Lee, M. Verleysen. *Nonlinear Dimensionality Reduction*, Information Science and Statistics, Springer, 2007.

[10] I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157-1182, 2003.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[11] J.G. Dy, C.E. Brodley, Feature subset selection and order identification for unsupervised learning. In *proceedings of the 17th International Conference on Machine Learning* (ICML 2000), Morgan Kaufmann Publishers Inc., pages 247-254, Standford, CA (USA), 2000.

[12] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications.* Studies in Fuzziness and Soft Computing, Springer, 2006.

[13] I.T. Jolliffe, *Principal Component Analysis.* Springer; 2nd edition, 2002.

[14] A. Vellido, E. Romero, M. Julià-Sapé, C. Majós, À Moreno-Torres, C. Arús, Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel proton MRS. *NMR in Biomedicine.* Accepted for publication. doi: 10.1002/nbm.1797.

[15] P.J.G. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*, 15(1):11-39.

[16] J. Venna, *Dimensionality Reduction for Visual Exploration of Similarity Structures.* Ph.D thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D20, Espoo, Finland, 2007.

[17] A. Wismueller, M. Verleysen, M. Aupetit, J.A. Lee, Recent advances in nonlinear dimensionality reduction, manifold and topological learning. In M. Verleysen, editor, *proceedings of the 18th European Symposium on Artificial Neural Networks* (ESANN 2010), d-side pub., pages 71-80, Bruges (Belgium), 2010.

[18] F. Camastra, A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1404-1407, 2002.

[19] T. Villmann, J.C. Claussen, Magnification control in Self-Organizing Maps and Neural Gas. *Neural Computation* 18(2):446-469, 2006.

[20] B. Hammer, A. Hasenfuss, T. Villmann, Magnification control for batch neural gas. *Neurocomputing* 70(7-9):1225-1234, 2007.

[21] A. Vellido, J.D. Martín, F. Rossi and P.J.G. Lisboa, Seeing is believing: The importance of visualization in real-world machine learning applications. In M. Verleysen, editor, *proceedings of the* 19th *European Symposium on Artificial Neural Networks* (ESANN 2011), d-side pub., pages 219-226, Bruges (Belgium), 2011.

[22] R. Miikkulainen, J.A. Bednar, Y. Choe, and J. Sirosh. *Computational Maps in the Visual Cortex*, Springer, 2005.

[23] H. Jeanny. *Vision: Images, Signals and Neural Networks. Models of Neural Processing in Visual Perception*, World Scientific Publishing, 2010.

[24] T. Kohonen. *Self-Organizing Maps*, (3rd ed.) Information Science Series, Springer, 2000.

[25] P.J.G. Lisboa, A. Vellido, B. Edisbury (Eds.) *Business Applications of Neural Networks.* Singapore: World Scientific, 2000.

[26] M. Aupetit, Visualizing distortions and recovering topology in continuous projection techniques, *Neurocomputing*, 70(7-9):1304-1330, 2007.

[27] Q. Du, V. Faber and M. Gunzburger, Centroidal Voronoi tessellations: Applications and algorithms, *SIAM Review*, 41(4):637-676, SIAM, 1999.

[28] V. Van Belle, S. Van Huffel, J. Suykens, S. Boyd, Interval coded scoring systems for survival analysis. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[29] J.A. Lee, M. Verleysen, Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431-1443, 2009.

[30] B. Mokbel, W. Lueks, A. Gisbrecht, M. Biehl, B. Hammer, Visualizing the quality of dimensionality reduction. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[31] T. Villmann, E. Merényi, W.H. Farrand, Unmixing hyperspectral images with Fuzzy Supervised Self-Organizing Maps. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[32] C.M. Bishop, M. Svensén and C.K.I. Williams, Magnification factors for the SOM and GTM algorithms. In *proceedings of the Workshop on Self-Organizing Maps* (WSOM'97), pages 333-338, June 4-6, Helsinki (Finland), 1997.

[33] A. Tosi, A. Vellido, Cartogram representation of the batch-SOM magnification factor. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[34] M.T. Gastner, M.E.J. Newman, Diffusion-based method for producing density-equalizing maps, *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499-7504, National Academy of Sciences, 2004.

[35] T. Rögnvaldsson, T.A. Etchells, L. You, D. Garwicz, I.H. Jarman, P.J.G. Lisboa, How to find simple and accurate rules for viral protease cleavage specificities. *BMC Bioinformatics*, 10:149, 2009.

[36] J.M. Martínez, P. Escandell, E. Soria, J.D. Martín, J. Gómez, J. Vila, Extended visualization method for classification trees. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[37] J. M Martínez, P. Escandell, E. Soria, J. D Martín, J. Gómez and J. Vila. Growing Hierarchical Sectors on Sectors. In M. Verleysen, editor, *proceedings of the* 19th *European Symposium on Artificial Neural Networks* (ESANN 2011), d-side pub., pages 239-244, Bruges (Belgium), 2011.

[38] L. Belanche, J. Hernández, Similarity networks for heterogeneous data. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[39] M. Kästner, W. Hermann, T. Villmann, Integration of structural expert knowledge about classes for classiffication using the Fuzzy Supervised Neural Gas. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[40] G. Zycinski, M. Squillario, A. Barla, T. Sanavia, A. Verri, B. Di Camillo, Discriminant functional gene groups identification with machine learning and prior knowledge. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[41] H. Ruiz, S. Ortega, I.H. Jarman, J.D. Martín, P.J.G. Lisboa, Constructing similarity networks using the Fisher information metric. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., current volume, Bruges (Belgium), 2012.

[42] C.H.Q. Ding, L. Tao, M.I. Jordan, Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45-55, 2010.

[43] H. Ruiz, S. Ortega, I. Jarman, A. Vellido, J.D. Martin, P.J.G Lisboa, Towards interpretable classifiers with blind signal separation. In *International Joint Conference on Artificial Neural Networks* (IJCNN 2012), Accepted for publication, 2012.

[44] G. Borboudakis, S. Triantafillou, V. Lagani, I. Tsamardinos, A constraint-based approach to incorporate prior knowledge in causal models. In M. Verleysen, editor, *proceedings of the 20th European Symposium on Artificial Neural Networks* (ESANN 2011), d-side pub., pages 321-326, Bruges (Belgium), 2011.

[45] P.J.G. Lisboa, T.A. Etchells, I.H. Jarman, C.T.C. Arsene, M.S.H. Aung, A. Eleuteri, A.F.G. Taktak, F. Ambrogi, P. Boracchi, E.M. Biganzoli, Partial Logistic Artificial Neural Network for competing risks regularised with Automatic Relevance Determination. *IEEE Transactions on Neural Networks*, 20(9):1403-1416, 2009.