ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Research directions in interpretable machine learning models

Vanya Van Belle[1,2] and Paulo Lisboa[2]

1- Department of Electrical Engineering (ESAT-SCD), KU Leuven/iMinds Future Health Department, Kasteelpark Arenberg 10/2446, 3001 Leuven, Belgium
2- School of Computing and Mathematical Sciences - Dept of Mathematics and Statistics Liverpool John Moores University, Byrom Street, Liverpol L3 3AF, UK

Abstract. The theoretical novelty of many machine learning methods leading to high performing algorithms has been substantial. However, the black-box nature of much of this body of work has meant that the models are difficult to interpret, with the consequence that the significant developments in machine learning theory are not matched by their practical impact. This tutorial stresses the need for interpretation and outlines the current status and future directions of interpretability in machine learning models.

1 Why interpretation and visualization in machine learning?

The above question directly corresponds in many applications to asking – why should machine learning methods be useful in practice? While there are many publications in this huge and significant field of learning, real-world applications are much fewer, especially in safety-critical domains. What are the reasons for this? How can flexible non-linear models be interpreted? Alternatively, given that there are different ways of articulating a flexible regression or classification model, can machine learning models be designed so that they are directly interpretable by construction? Is interpretation in effect an alternative view of Occam's razor, leading to models that not only generate better insights about the data but also better predictions for new data?

The reasons limiting the practical application of flexible models are several. Data can be limited and noisy, so generating marginal performance improvements at the cost of significant additional complexity – this is often the case with behavioral or biomedical data. Moreover, what is required is often an actionable model, not just estimates of conditional density functions. Therefore, notwithstanding the data deluge, making models useful requires that they generate insights about the data, not just abstract numerical predictions.

In some domains these requirements are enshrined in legal doctrines, such as the learned intermediaries: the liability for medical decision making rests with the physician, a principle that has been tested in court [1]. This hinges on the fundamental principle that the doctor is in the best position to assess risks and benefits. Put differently, "it is assumed that the physician has the expertise to understand, react to, and determine whether to override the clinical decision support system recommendation" [2].

This is in fact nothing more than considering the machine learning model to be part of the overall computer-based decision support system which for safety-critical applications needs to undergo hazards and operability analysis (HAZOP) to ensure,

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

for instance, safety of operation not just for the data used in training and validation but for all possible input values. It is a necessary step to enable the model to be completely verified by domain experts.

So now we have two requirements of interpretability. One is to generate insights about the data in such a way that the domain expert can validate the operation of the machine learning model against their expertise, so as to act as a learned intermediary between model and the actions derived from the model; the second is to verify that the model will remain valid across the full range of inputs – in machine learning terms, will not extrapolate beyond the training data. The second requirement is about novelty detection.

Moreover, data are often well-known to be non-linear. In these cases the need to be able to interpret the model at some level, frequently with a hybrid approach where the model remains linear-in-the-parameters, typically in the form of a General Linear Model, but non-linear data representations are used for instance by binning inputs into discrete intervals. This results in piecewise linear models that can have severe weaknesses, chief among them potentially gross errors that arise when even experience operators differ in the mapping of continuous values into bins, especially near cut points [3]. In practice, the same histological slide being submitted to different laboratories may be reported with different prognosis because of the significant effect of reporting one interval, e.g. pathological grade, rather than a neighboring grade – even though this is due to minor differences between subjective assessments near interval boundaries.

Therefore we can add a third requirement, namely that flexible models must overcome the limitations of linear-in-parameters models that require quantization of continuous values into discrete bands.

Taken together, these considerations serve to ascertain the level of interpretability and potential utility of flexible non-linear models, in summary, to:

1. Map onto domain knowledge.
2. Ensure safe operation across the full operational range of model inputs.
3. Accurately model non-linear effects.

These attributes reflect the intent in machine learning models of consolidating and analyzing information at the decision point, rather than supplanting the expertise of the practitioner. This level of transparency provides the user with legal protection [4] as well as adding valuable external consistency checks during the validation phase of the software development lifecycle [5].

In this tutorial, the current state of interpretable machine learning models and current directions will be outlined by reference to these three desirable factors, with only indicative references.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

2 Existing data mining tools with interpretable and/or visualisable results

Applicable models need to be accepted by practitioners and this has been articulated as requiring explicit documentation of the model structure, effects of uncertainty and limitations for its intended applications [4].

2.1 Nomograms

There are four broad categories of implementation of interpretable models that are generally used. The most traditional is the use of nomograms [6]. This is a natural description of generalized non-linear models since, by construction, these models apply a score function that is linear in the parameters. A familiar example of this is the linear term in logistic regression. Nomograms provide insight into the contribution of each covariate i.e. risk factor into the total risk score, so aiding the diagnostic process more than just through a one-to-one relationship with a posterior probability of class membership. However, this approach fails to meet the second and third desirable attributes since it neither verifies the applicability of the score index to new data, which may be from an outlier, and requires binning of continuous variables in order to model non-linear effects, as noted earlier.

An intuitive visualization of the weight of evidence i.e. the contribution of covariates in a risk model is a very attractive feature for interpretation of the model predictions. Retaining this for non-linear response surfaces, however, is a significant challenge. In this regard, progress has been made in binary classification and survival analysis. The Interval Coded Score model replaces implicit non-linear modeling by means of a kernel, so using an explicit parametric approach [7]. Each covariate is represented by a large number of binary indicators, each corresponding to an interval of covariate values. A sparsity mechanism is included to reduce the number of actually used binary indicators, as such representing an automatically defined score system. Interpretability of the obtained model is improved by providing a color based representation, where it is instantly noted that a red color tends to a bad prognosis/ diagnosis, and blue tends to a more positive prognosis/diagnosis. An example is given in Figure 1.

The extension of score indices to flexible models can take account of non-linearities through the use of kernels. A difficulty with this approach can be that the risk score is not well calibrated, for instance in the case of the standard SVM, where there is no direct correspondence between separation distances to the boundary and estimates of the posterior probability of class membership. This has the important consequence that while kernel models can be accurate in discrimination with hard binary decisions, they do not provide accurate estimates of the uncertainty in the predictions. However, progress has been made in linking discriminant methods from computational learning theory with probabilistic models [8].

From an interpretational viewpoint, the key factor is to explicitly identify interactions between covariates. This is best done by recourse to model sparseness, using the properties of discriminant models to select only the most informative terms in the kernel. A possibility is to restrict the form of the kernel to additive kernels, or ANOVA kernels that are able to incorporate interaction effects between covariates.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

A disadvantage of this approach is that prior knowledge is necessary in order to discard terms without a significant contribution. More recently, sparsity mechanisms were used in combination with a restricted Taylor expansion of the RBF kernel to select relevant main terms and two-way interaction terms of the covariates. As such, different contributions to the final score can be visualized as 2D maps (see Figure 2) [9].

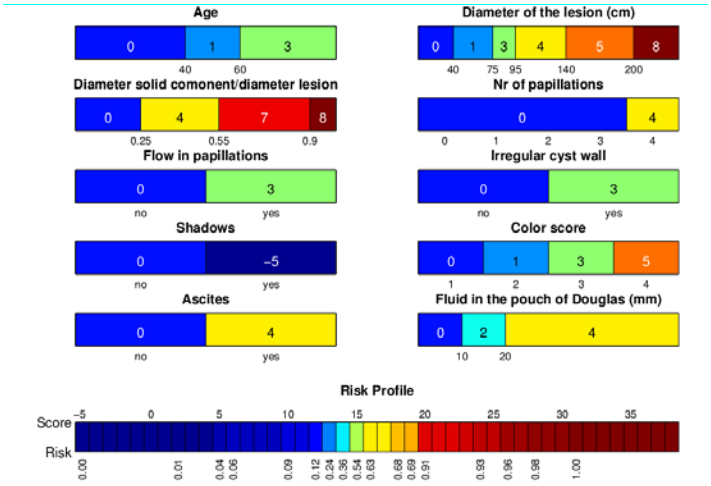

Fig1. Color based representation of a score system to calculate the risk on malignant ovarian tumors. [7]

Two ways to create interpretability by the use of a sparsity mechanism were discussed above. Both of them are used to perform a type of feature selection. However, sparsity can be used to increase interpretability in different settings. When using sparsity constraints in the dual setting, it is possible to select observations that can be used to represent the data, and thus can be interpreted as prototypes.

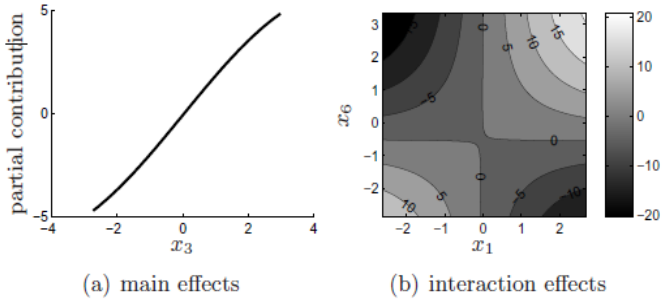

Fig. 2: Illustration of (a) main and (b) two-way interaction effects using a restricted Taylor expansion of the RBF kernel in combination with sparsity constraints. [9]

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

2.2 **Rule Induction**

The second most common interpretation of non-linear models is the induction of rule trees. Familiar examples include CART and MARS [10] which applies to regression as well as classification. Dendrograms are especially prevalent in bioinformatics since they cluster observations with hierarchical trees and so generate insights into the data structure. A significant limitation of many rule induction methods is reliance on sequential univariate decision points. This generates naturally orthogonal groups i.e. splitting the data into non-overlapping cohorts, but can be blind to important non-linear interactions between covariates unless multiple cut points are allowed for each covariate, with a consequent proliferation of rules. In addition, the results from rule induction can change even with small perturbations to the data. If this is the case, it seems the interpretation is not as straightforward as it seems. One way to attempt to stabilize the rules is through direct estimation of a discriminant model that may use kernels [11] or a probabilistic estimator of the probability of class membership, typically a heavily regularized neural network.

This alternative approach to rule extraction generates multivariate decision points in Disjunctive Normal Form with efficient search methods e.g. Orthogonal Search Rule Extraction (OSRE) [12] (see Figure 3). This generally facilitates interpretation by returning only a small number of low-order rules [13] albeit at the cost of orthogonality i.e. the rules now overlap. This means that the rules need to be sorted by their importance usually defined by the coverage of the data, i.e. sensitivity, and positive predictive value, which reflects specificity and class imbalance, or prevalence. A further advantage of this methodology is that it naturally 'boxes' the data into closed rules which together form a convex hull of the data density function used for rule estimation. Consequently, new cases that are outliers will typically fit none of the rules, so flagging for extrapolation.

Boolean rules are of course steep in the change in inference from one rule to the next. However step-changes are easily smoothed out with fuzzy logic, where overlapping intervals in the input data lead to partial membership of consequent rules forming an effective implementation of multi-linear interpolation.

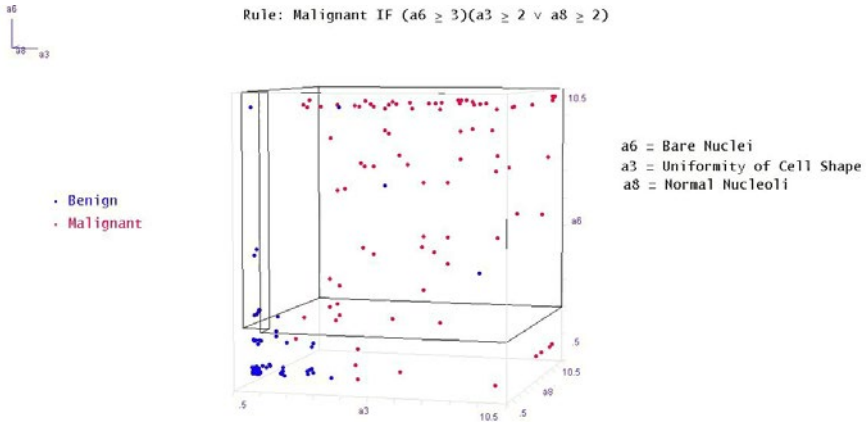

Fig. 3: Example of a rule generated by OSRE for the Wisconsin breast cancer dataset in the UCI repository [12].

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

2.3 Graphical Models

The third 'canonical' approach is in fact the most natural interpretation of the requirement to open-up the structure of the model to the domain expert. This is most readily done with conditional independence models, also called Bayesian Belief Networks [10], which represent visually a factorization of the joint density function of the data in terms of parent-child associations, shown by edges in the graph. The limitation of these models is that structure finding is NP-complete, therefore scalability to large data sets is in general low. The interpretation of the graphs can also suffer both from spurious nodes, i.e. false positives typically arising because of the need for repeated significance tests for conditional independence. However False Detection Rate (FDR) control is now widely used and this has been included in some scalable models [14] (see Figure 4). A limitation of this methodology is the need to find suitable parametric models for continuous variables.

An important development of graphical models is the direct estimation of causality [15] and the incorporation of prior knowledge into causal models [16].

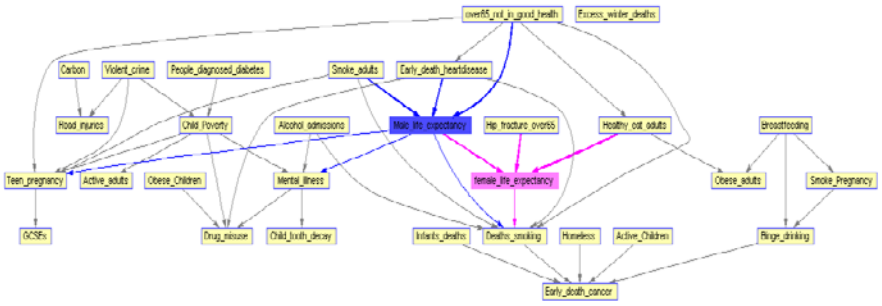

Fig. 4: Example of a Conditional Independence Map for life expectancy using data from UK Local Authorities [17].

2.4 Data Visualisation

Rather than describing the model structure to the user, a fourth approach is possible to show instead the structure of the data. This links interpretation with another complex domain in machine learning, data visualization, which is equally fraught with difficulties in deciding for instance how to evaluate them for objective comparisons between different methods. Nevertheless machine learning models are data intensive therefore understanding the data can be as crucial as understanding the model. Approaches to data structure range from vector quantization, including determination of prototypes[18] to topographic maps [19]. In a broad sense, direct querying of the data encompasses a broad range of apparently unrelated methodologies including also case-based reasoning and k-Nearest Neighbor models. A comprehensive review of this broad field would merit a full paper of its own.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

We will focus on a recent development that may be grouped under the term retrieval based classification. This is similar in concept to k-NN but instead of allocating a new data point to a new class, the purpose of the method is to identify the most relevant reference cases for the expert user to make this allocation [20]. In a sense this approach goes to the core of what the practitioner is trying to do. This is done by providing an objective tool based on the central concept of statistical similarity with respect to the classification. These methods have potential also for use as semi-supervised classifiers [21].

A principled approach to constructing a similarity measure from an estimate of a class conditional density function, namely the posterior probability of class membership *P(Class|x)*, is obtained from the Fisher Information Matrix. Having estimated this probability using a linear or flexible model, the Riemannian metric defined by the Fisher Information can be used to calculate pairwise geodesic distances which, in turn, maps the data into a network into which new data points can be integrated. This is the Fisher Information Network, from which communities of points can also be identified each with central nodes that may be suitable prototypes to describe the community [22] (see Figure 5). This approach is related to choices of kernel and also to feature extraction, since the Fisher Information metric naturally weights-up the most locally informative directions in data space, suppressing the rest.

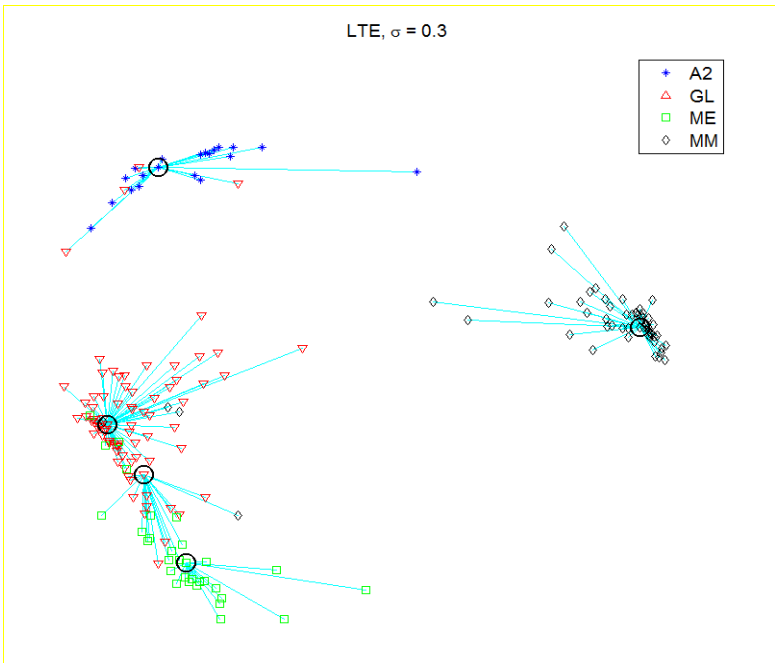

Fig. 5: Fisher Information Network for Magnetic Resonance Spectra from four different tumour groups using the methodology detailed in [22]. The figure shows network communities each with a prototype, of which there are five: one for each class and one for a mixture between two classes.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

3 Recent developments in sparse and interpretable models

Having established four generic approaches to interpretation of machine learning systems, we now turn to emerging research directions in this field, which are represented by papers in the current Special Session on *Sparsity for interpretation and visualization in inference models.* Following last year's special session [23] the aim is now to embed interpretation into the design of machine learning systems by constructing them to be parsimonious, on the hypothesis that the simpler the model so more interpretable it will be.

The first paper puts this intuition into practice by modelling regression with only the main effects, expressed here in symbolic form. This is often what is done in practice, but the novelty in the paper is to formally assess the detriment in performance arising from the simplification of the model using a flexible model of the modeling error to obtained guaranteed bounds for inferences on new data.

The second paper takes a completely different approach to sparseness, enforcing this through kernels. This leads to the identification of prototypes in the context of Learning vector Quantization, which also provides an estimate of the convex hull of the data.

The third paper focuses on the identification and visualization of outliers through non-linear modeling of the data density function, expressed visually in a cartogram representation showing the magnification factors for a generative topographic map with robust statistics.

Following the oral sessions there will be three further presentations on visualization methodologies to show non-linear dependencies in high-dimensional data, demonstrated in practical applications.

4 Acknowledgments

The authors are grateful to I.H.Jarman, T.A. Etchells and H. Ruìz for helpful discussions and production of figures for this paper. This research is supported by GOA MaNet, PFV/10/002 (OPTEC), G.0108.11 (Compressed Sensing), iMinds, IUAP P7/ (DYSCO, 'Dynamical systems, control and optimization',2012-2017), ERC AdG A-DATADRIVE-B. VVB is a postdoctoral fellow of the Research Foundation – Flanders (FWO).

References

[1] www.thedoctors.com/KnowledgeCenter/Publications/TheDoctorsAdvocate/CON_ID_002961, accessed on 18/2/2013.

[2] E. S. Berner, Ethical and legal issues in the use of clinical decision support systems, *Journal of Healthcare Information Management*, 16(4):34-37, 2002.

[3] C. Bennette and A. Vickers BMC, Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents, *Medical Research Methodology*, 12:21, 2012.

[4] S. R. Klein and J. W. Jones, Clinical decision support programs can be risky business. *J Healthc Inf Manag.*, 21(2):15-7, 2007.

[5] D. M. Eddy, W. Hollingworth, J.J. Caro, et al. Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices task force-4, *Value Health,*15:843-50, 2012.

[6] M. Kattan and J. Marasco, What Is a Real Nomogram?, *Seminars in Oncology*, 37(1): 23–26, 2010.

[7] V. Van Belle, B. Van Calster, D. Timmerman, T. Bourne, C. Bottomley, L. Valentin, P. Neven, S. Van Huffel, J.A.K. Suykens and S. Boyd, A Mathematical Model for Interpretable Clinical Decision Support with Applications in Gynecology', *PLoS ONE*, 7(3): 1-10., 2012.

[8] H. Chen, P. Tiňo and X. Yao, Probabilistic Classification Vector Machines, IEEE Trans Neural Networks, 20(6):901-914, 2009.

[9] V. van Belle, S. van Huffel, J. Suykens and S. Boyd, Interval coded scoring systems for survival analysis, In ESANN'12, pages 173-178. d-side pub., 2012.

[10] T. Villmann, F.-M. Schleif and B. Hammer, Spare representations of data, In ESANN'10, pages 225-234.d-side pub., 2010.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[12] Z. Chen, J. Li and L. Wei, A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artif Intell Med*, 41(2):161-75, 2007.

[13] T. A. Etchells and P. J. G. Lisboa, Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach, *IEEE Transactions on Neural Networks*, 17 (2): 374–384, 2006.

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

[15] D. Bacciu, T. A. Etchells, P. J. G. Lisboa and J. Whittaker, Efficient identification of independence networks using mutual information, *Computational Statistics* DOI 10.1007/s00180-012-0320-6, 2012.

[16] T. Claassen and T. Heskes. A structure independent algorithm for causal discovery. In ESANN'11, pages 309-314.d-side pub., 2011.

[17] G. Borboudakis, S. Triantafilou, V. Lagani and I. Tsamardinos A constraint-based approach to incorporate prior knowledge in causal models. In ESANN'11, pages 303-308.d-side pub., 2011.

[18] H. Carlin, I. H. Jarman, S. Chambers, P. J. G. Lisboa, S. Knuckey, C. Perkins and M. A. Bellis 'North West Mental Wellbeing Survey – what influences wellbeing?', NHS North West Commissioned Report by the North West Public Health Observatory, May 2011.

[19] T. Villmann, F.-M. Schleif and B. Hammer, Spare representations of data, In ESANN'10, pages 225-234.d-side pub., 2010.

[20] 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 ESANN'11, pages 219–226.d-side pub., 2011.

[21] G. Rendes and A. Tsymbal, HeC CaseReasoner: Neighborhood Graph for Clinical Case Retrieval and Decision Support, Proc. LWA 2009 Darmstadt, 21-23 September, 2009.

[22] R. G. F. Soares, H. Chen and X. Yao, Semisupervised classification with cluster regularization, IEEE Tran *Neural Networks and Learning Systems*, 23(11):1779-1792, 2012.

[23] H. Ruiz, T. A. Etchells I. H. Jarman, J. D. Martín, and P. J. G. Lisboa, A principled approach to network-based classification and data representation, *accepted for Neurocomputing*, 2013.

[24] A. Vellido, J. D. Martín, and P. J. G. Lisboa. Making machine learning models interpretable. In ESANN'12, pages 163–172. d-side pub., 2012.