Seeing is believing: The importance of visualization in real-world machine learning applications

Alfredo Vellido[1], José D. Martín[2], Fabrice Rossi[3] and Paulo J.G. Lisboa[4] *

1- Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya, 08034, Barcelona - Spain

2- Departament d'Enginyeria Electrònica, Escola Tècnica Superior d'Enginyeria
University of Valencia, 46100, Burjassot(Valencia) - Spain

3- Computer Science & Networks Department
Télécom ParisTech, F-75634, Paris - France

4- Department of Mathematics and Statistics
Liverpool John Moores University, Byrom Street, L3 3AF, Liverpool - UK

Abstract. The increasing availability of data sets with a huge amount of information, coded in many different features, justifies the research on new methods of knowledge extraction: the great challenge is the translation of the raw data into useful information that can be used to improve decision-making processes, detect relevant profiles, find out relationships among features, etc. It is undoubtedly true that a picture is worth a thousand words, what makes visualization methods be likely the most appealing and one of the most relevant kinds of knowledge extration methods. At ESANN 2011, the special session "Seeing is believing: The importance of visualization in real-world machine learning applications" reflects some of the main emerging topics in the field. This tutorial prefaces the session, summarizing some of its contributions, while also providing some clues to the current state and the near future of visualization methods within the framework of Machine Learning.

1 Introduction

Data exploration is one of the basic building blocks, or constituting stages, of most standard Data Mining (DM) [1, 2] and Knowledge Discovery in Databases (KDD) [3] methodologies, either on its own or as part of a more generic phase of data understanding. It could be defined as the use of techniques, from data querying and basic statistics to advanced visualization, to discover the main characteristics of usually complex multivariate data sets, helping bring important aspects of the data into focus for study in subsequent phases of the analysis.

The task of data visualization is central to data exploration. So central, in fact, that there would be arguments in favor of considering it a DM phase on its own. Note though that even the consideration of data exploration as an independent DM phase has been found to be controversial[1].

*This research was partially supported by Spanish MICINN projects TIN2009-13895-C02-01, TIN 2007-61006 and CSD2007-00018.

[1] http://crispdm.wordpress.com/2007/03/13/exploration-can-anyone-out-there-explain-it/

The problem of knowledge generation through information visualization [4], as a generalization of that of data visualization, is not circumscribed to the fields of DM and KDD. It has in fact been a matter of study in other areas [5] and it could be argued that it is a type of problem to be addressed from the viewpoints of both artificial pattern recognition (APR) and natural pattern recognition (NPR): the former through the definition of visualization-oriented techniques stemming from the fields of artificial intelligence and statistics; the latter through the understanding of visualization as the cognitive processing of visual stimuli conducted by the human brain (which can indeed be investigated using machine learning techniques [6, 7] in the context of Computational Neuroscience [8]).

Humans are equipped with visual NPR as a tool to understand the patterns of their natural environment and operate upon it. As Lehrer [9] nicely puts it while arguing on the influence of imaging techniques in the current development of neuroscience, adequate data visualizations can help us to gain insights into a problem "without the frame of a conjecture". They do so by letting us, even if temporarily, slip out of a deductive model of research to reap the benefits of a more inductive one. This may be crucial when modeling the complex multivariate and heterogeneous data that are increasingly common in many areas of research. Importantly for us, as data analysts, APR and NPR can enhance each other in order to make data exploration a more fruitful process. As stated in [10], integrated processes that unify DM algorithms and visual user interfaces should allow us to explore data using graphical metaphors, in a way that helps to circumvent some of the inherent limitations of human vision [11, 12].

In order to explain the success of visualization as a data exploration tool, we should not brush aside the aesthetic aspects involved. Put simply, pretty pictures are useful because they appeal to us at a very basic, non-discursive level. Usefulness and beauty can certainly be indistinguishable concepts in this context: as argued in [13], for a data visualization to qualify as beautiful, it must ultimately comply with requirements of novelty, informativeness, and efficiency.

This brief tutorial does not aim to cover the issue of data visualization in full. Instead, it delves into a particular area, namely health, in which the potential of data visualization is illustrated in some detail. Readers are then provided with some general outlook of the main opportunities and challenges concerning the use of computational intelligence approaches in data visualization. Finally, the main contributions of the papers accepted for the ESANN 2011 special session that this tutorial prefaces are summarily discussed.

2 Data Visualization in Health

Visualization is absolutely central to the communication of complex information in a way that is rapidly absorbed and conveys the necessary insight. We are all familiar with the astounding progress in medical imaging, which started as a window to anatomical structures, but is increasingly driven by functional methods which show metabolic activity such as, for instance, glucose metabolism. This field is moving towards the use of detailed 3D maps of the concentration of

specific metabolites, entering the era of molecular imaging. An example that combines signal processing for tissue segmentation using Magnetic Resonance Spectroscopy (MRS) overlaid onto a high resolution anatomical image can be found in [26].

Clearly, direct rendering of physiological activity is as close as we can get to the locus of disease. More generally, the role of visualization is the last cognitive step in intelligent data analysis, linking individual observations to the structure of the rest of the data set. This involves mapping as much of the data as possible into a low dimensional projection, while retaining the proximity structure and with as little distortion as possible. Given the clear preference in health for linear methods, canonical analysis remains the projection method of choice [27]. Besides Fisher Linear Discriminant Analysis, linearly separable cohorts arising e.g. from clustering may be projected still in linear space while achieving significant dimensionality reduction with minimal, or no mixing, of the cohorts [28]. This approach to low-dimensional visualization with scatter matrices has since been developed for diagnostic classification of MRS [29].

Beyond the application of linear methods, lies what we normally term computational intelligence methods. These methods comprise three broad categories of which only the first is well developed. They are non-linear projections, directed graphs and proximity networks.

Non-linear projections with neural networks hark back to the Self-Organizing Map, which has been developed to project high-dimensional, time-varying information, in 2-D maps that correlate with diagnostic features, as in [30]. The further development of probabilistic non-linear dimensionality reduction techniques has also included different data density functions [31], including kernels [32]. There are, of course, many other methods to map data structure with non-linear models, mostly used in exploratory data analysis rather than intended for direct human-computer interfaces.

While projective methods focus on the visual display of the structure of labeled observations, commonly represented as the rows in the data matrix, it is often as important to visualize the relationship between the covariates i.e. the associative structure of the row elements. This is particularly the case in bioinformatics for the discovery of activity pathways linking the genotype with the disease expression, or phenotype. An example of the use of graphs in the context of data mining is outlined in [33]. These methods are central for the elucidation of functional structures that underpin the deep mechanisms of pathogenesis.

Network visualization and structural analysis is a fast growing area of research with applications in many fields including bioinformatics. In particular, this approach is leading to re-categorization of disease sub-types on the basis of molecular information, an example of which can be found in [34]. In this respect, they form a unifying link between molecular imaging, deeper understanding of expression pathways, and graphical models which may be used to clean noise in association maps.

In summary, visualization has grown to encompass projections of the geometric distribution of data points, usually to show the proximity between rows in the

data matrix, but also becoming an integral part of the methodology actively involved in unlocking networks of functional relationships between covariates, from which to derive deep insights into the mechanisms driving disease processes.

3 Computational Intelligence in Data Visualization

As argued in the introduction and illustrated in the previous section, data visualization and computational intelligence are tightly linked: many DM methods, especially unsupervised ones, rely on human control and monitoring enabled through visual tools and reports, while many visualization techniques are built upon DM algorithms. For instance, visual exploration of the clustering structure of a high dimensional dataset can be done using a nonlinear dimensionality reduction method [14] combined with a standard scatter plot augmented with dynamic distortion [15] and with visualization of projection errors [16]. In other words, in order to help a future clustering algorithm, one relies on sophisticated visualization methods that are in turn enabled by an automated manifold learning method (or any other dimensional reduction technique).

One of the main roles of DM algorithms in visualization methods consists in providing some form of scalability [11]: the human vision is strongly limited to 2D/3D displays, with the additional ability to decode shapes and colors efficiently. Thus, no more that 5 to 6 variables can be displayed at a time, provided the number of data points remains limited, in order to avoid overlapping. Then, feature selection [17], dimensionality reduction [14] and clustering [18] are methods of choice as preprocessing solutions for medium to large scale data visualization.

However, it remains difficult to match DM methods with visualization tools, mainly because quality criteria used to implement e.g. dimensionality reduction are generally unrelated to visual qualities of the display. Some progress has been made in this direction by designing cases-oriented quality measures: in the case of clustering analysis, for instance, it is important for neighborhood structures on the visual representation to reflect accurately their high dimensional counterpart and this can be assessed via rank comparisons [19]. Unfortunately, it is well known that such quality measures are generally both very difficult to optimize directly (because of their combinatorial nature) and non-consensual. For instance, while minimizing the number of edge crossings is considered as one of the most important quality criteria for graph drawing [20], many other criteria have been shown to have influence on the way a graph is perceived [21]; moreover, the simple fact of drawing edges of a graph modifies to a large extent the way distances between nodes are perceived [22].

Additionally, the interplay between visualization and DM is still minimal. Some contributions stem from the use of the Self-Organizing Map (SOM) [23], one of the earliest successful combinations of computational intelligence with visualization methods, extended to more recent models such as the Generative Topographic Mapping (GTM) [24] and its interactive hierarchical variant [25]. Despite this progress, further work will be required to actually reach the visual DM goal, namely to provide visualizations that are fully integrated with DM, in a

way that permits, for instance, user feedback to be re-injected into DM methods so as to improve the visual results on the fly, while minimizing misinterpretation risks.

Papers of this session show examples of successful hybridisation between visualization methods and DM: [35] provides a new visualization method for a DM method (the growing hierarchical SOM), [36] leverages clustering to simplify complex networks prior drawing them, [37] uses the SOM to ease meta-parameter tuning for DM methods, and [38] uses a kernel GTM to visualization complex non vector data. The following Section details their contributions.

4 Seeing is believing: Data Visualization in ESANN 2011

Four contributions were accepted to the special session "Seeing is believing: the importance of visualization in real-world machine learning applications" at ESANN 2011. They address diverse theoretical and application issues which are organized in two main themes, glossed next.

4.1 Hierarchical visualizations

Visualization is one of the cornerstones of knowledge extraction from large databases. In this framework, hierarchical approaches appear as a natural solution since global methods producing a single "picture" of the data may provide either too complicated or too simplistic visualizations, as they may lack the detail crucial for data understanding and knowledge extraction. Hierarchical methods can produce visualizations at different levels of the hierarchy of detail, thus obtaining both main coarse relationships and detailed information, depending on the level of the hierarchy we focus on. This is studied in two contributions of this special session, [35] and [36].

In [35], a new visualization approach for the Growing Hierarchical Self-Organizing Map (GHSOM) model, a hierarchical variant of SOM, is proposed. Since the main limitation of GHSOM is that it is not possible visualizing simultaneously the data information at each level of the hierarchy, this paper presents a visualization method based on pie charts that does allow a simultaneous and compact visualization of the different hierarchy levels. The method is tested in synthetic and real data sets with internal hierarchical structure. The satisfactory performance achieved reinforces the viability of this method in hierarchical data visualization, since it enables the extraction of information by inferring relationships among features, neurons and levels of the hierarchy.

In [36], the proposed visualization makes use of clustered graphs based on hierarchical maximal modularity clustering. The general layout of clustered graphs is given by the higher level of the hierarchy in which the graph is strongly simplified, and then details are added in a top-down way by descending along the hierarchy. This paper presents a methodology to address two of the main limitations of clustered graphs, namely the way of obtaining the hierarchical clustering and its significance, which is usually not questioned despite the fact that a bad clustering solution may lead to interpretation errors. An efficient

modularity visualization is obtained by means of a variant of the Fruchterman-Reingold algorithm, which allows interactive exploration, with possibilities of coarsening and refining. The methodology is tested on a dataset that contains information about sexual contacts between patients infected by the HIV in Cuba.

4.2 Visualization in manifold learning

As expected in any session on the theme of visualization, manifold learning and, more specifically, SOM plays an important role in some of the contributions. Actually, two of contributions to this session [35, 37], propose approaches based on SOM variants. As previously explained, [35] proposes a visualization for GHSOM. The most explicit use of SOM can be found in [37], which resorts to SOM visualization to infer the relationships among the different parameters that need to be tuned in Machine Learning algorithms, thus helping to find the optimal combination of parameters that maximizes the performance of the algorithm. Although the approach presented in [37] is particularly applied to Reinforcement Learning, the strategy is completely general and thus applicable to any other kind of algortithm.

Visualization based in a kernel version of Generative Topographic Mapping (KGTM), which is a probabilistic reformulation of SOM, is presented in [38]. This work provides new insights in the study of G-protein-coupled receptors (GPCR's) sequences, which are a common target in pharmaceutical research as they regulate the function of most cells in living organisms. Sequence analysis can help in increasing the knowledge of the GPCR function. The main contribution of the paper is the exploitation of the probabilistic properties of KGTM to deal with non-quantitative information and to visually explore GPCR subclasses in detail, thus obtaining a map of probability that can qualify the differences between sequences belonging to different model groupings.

5 Conclusions

Hierarchical visualizations and manifold learning constitute two of the main Machine Learning approaches to produce visualizations that can extract knowledge from data sets. Although there has been intense research on these topics lately, work in the area is still very far away from being over. Available data sets get ever bigger, more complex, structured and heterogeneous. Therefore, there is a need for new methods that can extract knowledge from data sets in a straightforward way. This can have an automatic effect in real practice, since the appealing results produced by visualization methods are usually easily understandable by application experts who are not necessarily knowledgeable about Machine Learning. These experts may opt to choose this kind of methods rather than others, even if producing satisfactory results, may be of more difficult understanding.

References

[1] C. Shearer, The CRISP-DM model: The new blueprint for Data Mining, *Journal of Data Warehousing*, 5(4):13–22, 2000.

[2] G. Held, The process of data mining. In *proceedings of the Eighth International Symposium on Applied Stochastic Model and Data Analysis*, (ASMDA 1997), pages 155–164, June, Anacapri, Italy, 1997.

[3] U. Fayyad, G. Piatetski-Shapiro and P. Smith, From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 17(3):37–54, 1996.

[4] E.R. Tufte. *Envisioning Information*. Graphics Press, USA, 1990.

[5] J. Steele, N. Iliinsky, editors. *Beautiful Visualization*, O'Reilly Media, 2010.

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

[7] H. Jeanny. *Vision: Images, Signals And Neural Networks. Models Of Neural Processing In Visual Perception*, World Scientific Publishing, 2010.

[8] T.P. Trappenberg. *Fundamentals of Computational Neuroscience*, Oxford University Press, 2002.

[9] J. Lehrer, foreword to C. Schoonover, editor. *Portraits of the Mind: Visualizing the Brain from Antiquity to the* 21^{st} *Century*, Harry N. Abrams, Inc., 2010.

[10] J. Gray, foreword to U. Fayyad, A. Wierse, G.G. Grinstein, editors. *Information Visualization in Data Mining and Knowledge Discovery*, Academic Press, 2002.

[11] F. Rossi. Visual data mining and machine learning. In M. Verleysen, editor, *proceedings of the* 14^{th} *European Symposium on Artificial Neural Networks* (ESANN 2006), d-side pub., pages 251-264, April 26-28, Bruges (Belgium), 2006.

[12] R. Fuchs, J. Waser and M.E. Gröller, Visual human+machine learning, *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1327–1334, 2009.

[13] N. Iliinsky, On beauty. In J. Steele, N. Iliinsky, editors. *Beautiful Visualization*, O'Reilly Media, 2010.

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

[15] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction*, 1(2):126–160, 1994.

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

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

[18] A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering: a review, *ACM Computing Surveys*, 31(3):264–323, 1999.

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

[20] C. Ware, H.C. Purchase, L. Colpoys and M. McGill, Cognitive measurements of graph aesthetics, *Information Visualization*, 1(2):103–110, 2002.

[21] H.C. Purchase, Metrics for graph drawing aesthetics, *Journal of Visual Languages and Computing*, 13(5):501–516, 2002.

[22] S. Fabrikant, D. Montello, M. Ruocco, and R. Middleton. The distance-similarity metaphor in network-display spatializations. *Cartography and Geographic Information Science*, 31 (4):237–252, 2004.

[23] T. Kohonen. *Self-Organizing Maps*, Springer Series in Information Sciences, vol.30, Springer, 3^{rd} edition, 1995. Last edition published in 2001.

[24] C. M. Bishop, M. Svensén, and C. K. I. Williams, GTM: The generative topographic mapping, *Neural Computation*, 10(1):215–234, 1998.

[25] P. Tino and I. Nabney, Hierarchical GTM: Constructing localized non-linear projection manifolds in a principled way, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):639–656, 2002.

[26] J. Luts, T. Laudadio, A.J. Idema, A.W. Simonetti, A. Heerschap, D., Vandermeulen, J.A.K. Suykens and S. Van Huffel, Nosologic imaging of the brain: segmentation and classification using MRI and MRSI. *NMR in Biomedicine* 22(4):374-390, 2009.

[27] S. Ortega-Martorell, I. Olier, M. Julià-Sapé and C. Arús, SpectraClassifier 1.0: a user friendly, automated MRS-based classifier-development system, *BMC Bioinformatics*, 11(1):106, 2010.

[28] P.J.G. Lisboa, I.O. Ellis, A.R. Green, F. Ambrogi and M.B. Dias, Cluster-based visualisation with scatter matrices, *Pattern Recognition Letters*, 29(13):1814-1823, 2008.

[29] A. Vellido, E. Romero, F. González, L. Belanche, M. Julià-Sapé and C. Arús, Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database, *Neurocomputing*, 72(13-15):3085-3097, 2009.

[30] G.J. Barton, A. Lees, P.J.G. Lisboa and S. Attfield, Gait quality assessment using self-organising artificial neural networks, *Gait and Posture*, 25(3):374-379, 2007.

[31] A. Vellido and P.J.G. Lisboa, Handling outliers in brain tumour MRS data analysis through robust topographic mapping, *Computers in Biology and Medicine*, 36(10):1049-1063, 2006.

[32] I. Olier, A. Vellido, and J. Giraldo *Kernel Generative Topographic Mapping*. In M. Verleysen, editor, *proceedings of the* 18^{th} *European Symposium on Artificial Neural Networks* (ESANN 2010), d-side pub., pages 481-486, April 28-30, Bruges (Belgium), 2010.

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

[34] S.K. Bhavnani, F. Eichinger, S. Martini, P. Saxman, H.V. Jagadish and M. Kretzler, Network analysis of genes regulated in renal diseases: implications for a molecular-based classification, *BMC Bioinformatics*, 10(Suppl 9):S3, 2009.

[35] J. M. Martínez, P. Escandell, E. Soria, J.D. Martín, J. Gómez, J. Vila. Growing hierarchical sectors on sectors. In this volume, 2011.

[36] S. Clémençon, H. De Arazoza, F. Rossi, V. Tran. Hierarchical clustering for graph visualization. In this volume, 2011.

[37] V. Buendía, E. Soria, J.D. Martín, P. Escandell, J. M. Martínez. Analysis of a Reinforcement Learning algorithm using Self-Organizing Maps. In this volume, 2011.

[38] A. Vellido, M.I. Cárdenas, I. Olier, X. Rovira, J. Giraldo. A probabilistic approach to the visual exploration of G Protein-Coupled Receptor sequences. In this volume, 2011.