ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Information Visualisation and Machine Learning: Characteristics, Convergence and Perspective

Benoît Frénay and Bruno Dumas

Université de Namur - Faculty of Computer Science - PReCISE
Rue Grandgagnage 22, 5000 Namur - Belgium

Abstract. This paper discusses how information visualisation and machine learning can cross-fertilise. On the one hand, the user-centric field of information visualisation can help machine learning to better integrate users in the learning, assessment and interpretation processes. On the other hand, machine learning can provide powerful algorithms for clustering, dimensionality reduction, data cleansing, outlier detection, etc. Such inference tools are required to create efficient visualisations. This paper highlight opportunities to collaborate for experts in both fields.

1 Introduction

Information visualisation and machine learning originate from different fields. Whereas information visualisation is closely related to human-computer interaction, graphics and psychology, machine learning relies on concepts from applied mathematics, statistics, neurology, etc. However, they both aim to amplify human cognition for the visualisation and analysis of complex, massive data. Since these two tasks are complementary, both research fields are fated to cross-fertilise. On the one hand, information visualisation provides tools to put users back at the center of machine learning. For example, visualisation is necessary in model quality assessment, model prediction assessment and model examination. On the other hand, machine learning can be used to cleanse data and to extract knowledge that information visualisation would otherwise be unable to expose. This knowledge extraction can be performed with machine learning techniques such as clustering, dimensionality reduction or outlier detection.

Recently, many conferences, workshops and seminars have offered a place to discuss common issues and to share solutions. This paper advocates for such events and introduces the contributions of the special session "Information Visualisation and Machine Learning: Techniques, Validation and Integration" of the ESANN'16 conference. First, Sections 2 and 3 discuss the information visualisation and machine learning perspectives, respectively. Then, Section 4 shows how both research fields can integrate and converge to enrich each other.

2 The Information Visulisation Perpective

Information visualisation is the study of how to represent abstract data in a visual way to help provide insight and understanding of abstract data. This capability of visual representations to help humans get insight about data is also referred to as "visual data mining", as will be detailed further below. In [1],

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Card, Mackinlay and Shneiderman present this generally well accepted definition of information visualisation:

> "*Information visualisation is the use of computer-supported interactive, visual representation of abstract data to amplify cognition.*"

This rather packed definition describes four different aspects at the core of information visualisation. Let us go through these four different aspects.

First, the definition mentions [computer-supported] *visual representations*. Relying on the visual human capabilities is at the core of information visualisation. In particular, human beings have been shown to be very efficient at detecting trends or outliers in a visual representation, in particular through their preattentive processing capabilities [2]. For example, in a set of red-colored dots, a group of blue-colored dots will get spotted directly by the human low-level visual system. This is a task achieved non-consciously by human beings, and information visualisation relies in part on this powerful human capability.

A second core aspect of information visualisation is that it relies on *abstract data*. This contrasts with scientific visualisation, such as a weather forecast or a visual representation of the human brain. More generally, when we refer to abstract data in the domain of information visualisation, we refer to data which has no real-world "visual" counterpart (as would be the case with a visualisation of the different brain lobes), and for which there is no "natural" representation (as would be the case with the weather forecast).

A third core aspect of information visualisation lies in *interactivity*, that is, the possibility for the user to explore and manipulate the visual representation. Much of the potential power of a good visualisation relies on careful interaction [3], and it may be an aspect that still offers many promising venues at the frontier between machine learning and information visualisation. Schneiderman, in [4], presented the Information Seeking Mantra: "*overview first, zoom and filter, and then details-on-demand*". A classical illustration of this mantra is the way a user navigates an online map such as Google Maps. Numerous techniques have been introduced for enabling interactivity in visualisations, such as zooming and panning, semantic zooming or brushing and linking [5].

Finally, the last core aspect of information visualisation lies in its capability to *amplify cognition*. This is notably what is covered by the terms of visual data exploration and visual data mining. Three roles can thus be identified for information visualisation: a role of *exploration*, where a new hypothesis is formulated; a role of *confirmation*, where an existing hypothesis is confirmed or rejected; and a third role of *communication*, where a previously confirmed hypothesis is demonstrated. These three roles lead to the visual analytics' motto: "*detect the expected and discover the unexpected*". As Keim puts it [5], "visual data exploration [or visual data mining, or visual data analysis] is especially useful when little is known about the data and the exploration goals are vague".

The task of visual data mining thus typically involves the interactive visual exploration of massive datasets. Examples of such exploration may include cluster analysis, outlier detection, dependency assessment or pattern detection

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

(repetition, sub-structure, etc.). As one can see, these are typical problems that have been explored, among others, in the machine learning field. Machine learning approaches work especially well when confronted with well-defined questions, typically on very large and/or multidimensional datasets. In contrast, information visualisation approaches have shown their efficiency when a specific question has yet to emerge from a dataset. This is also one instance where machine learning and information visualisation approaches can complement each other.

This potential dialogue between human-centered visual data mining and machine learning approaches is one of the many promising venues at the intersection between these two worlds. Alongside mutual dialogue, both worlds have also the potential to greatly enrich each other. On the one hand, machine learning researchers have long relied on visual representations to get a visual assessment of the performances of their algorithms. On the other hand, information visualisation practitioners have had to fight since the beginning with noisy or huge datasets, frequently relying on ad-hoc filtering or clustering methods to produce readable, meaningful visualisations. In both cases, researchers from both fields applied techniques from the other. This article later reviews promising venues where both fields can actively contribute to each other.

3 The Machine Learning Perpective

Similarly to information visualisation, machine learning can also be seen as a set of tools to amplify cognition. Indeed, machine learning allows humans to find patterns in datasets that are too large to be grasped, to get insight in very complex processes, to predict trends in very fast time series, etc. Without machine learning tools, addressing such tasks would be a mere dream. However, visualisations are also essential to communicate machine learning results to users, whose cognition is to be amplified and who are often not machine learning experts. In this section, three cases are considered where visualisation is necessary: model quality assessment, model prediction assessment and model examination.

Model quality assessment is an important issue in machine learning, as discussed in [6]. In supervised classification, this is typically done using accuracy and similar metrics. Indeed, the goal of the user is clearly specified: maximising the amount of correct classifications (although, in some cases, class imbalance or misclassification costs should be taken into account). On the contrary, designing quality metrics in clustering and dimensionality reduction is much more subjective. One the one hand, there often exists no ground truth, except for artificial toy problems. On the other hand, clustering and dimensionality reduction are often considered as ill-defined problems. For example, there exist many definitions of *what a cluster is*, leading to a plethora of clustering methods. For such unsupervised problems, getting informative feedback from users is essential, what the user-centric field of information visualisation can help for.

Machine learning models are often used to make predictions about new, unseen instances. Assessing the quality of these predictions can be achieved using model quality metrics, but it is not sufficient. For example, in classification,

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

although accuracy is an objective measure of classifier quality, it does not tell anything about the predictions themselves. The user may be interested in visualising what the predictions look like, where the model makes mistakes, etc. This exploration task can benefit from information visualisation tools. Prediction assessment is particularly important in dimensionality reduction, since the goal is often to provide useful visual representations of high-dimensional data.

Model examination is closely connected to prediction visualisation. Indeed, model visualisations aim to provide insights on how a given model works and achieves its decisions. A typical example are decision trees that can easily be interpreted by non-experts. Since model examination is possible only if models can be interpreted by the user, many works have studied the problem of model interpretability, like e.g. [7, 8, 9]. Again, the user is crucial to assess model interpretability and several recent works use surveys [10, 11, 12] and metrics of user performances [13], what are typical tools in information visualisation.

As a final remark, notice that visualisations are also used in the machine learning literature to compare algorithms. For example, box plots can be used to show the difference in mean and deviation between the performance of several models. Recently, more complex tools have been proposed, like e.g. the Nemenyi and Bonferroni-Dunn tests [14] that allow comparisons over multiple datasets.

4 Towards Integration and Convergence

As described before, information visualisation and machine learning have almost from their respective beginnings relied on techniques borrowed from each other's field. Practitioners from both fields are more and more organising common venues to help discuss issues where the expertise developed from each side may help improve the other[6, 15].

Bertini et Lalanne [16] observed three different distinctive patterns of collaboration between the fields of information visualisation and machine learning (quoted text is from [16]):

1. *Computationally enhanced Visualization* "contains techniques which are fundamentally visual but contain some form of automatic computation to support the visualization"

2. *Visually enhanced Mining* "contains techniques in which automatic data mining algorithms are the primary data analysis means and visualization provides support in understanding and validating the result"

3. *Integrated Visualization and Mining* "contains techniques in which visualization and mining are integrated in a way that it is not possible to distinguish a predominant role of any of the two in the process."

Some interesting issues relevant to the *Computationally enhanced Visualization* category include machine learning-based techniques that would help ensure the readability of a visualisation. Typically, when creating visualisations of huge

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

datasets, most techniques result on cluttered, occluded, or downright unreadable representations, and information visualisation practitioners generally resort to processing or filtering the original data by hand. Generally speaking, scalability of visualisation techniques has been a long-standing issue in the field.

Regarding the *Visually enhanced Mining* category, Section 3 shows that visualisation tools are necessary for common tasks such as model quality assessment, model prediction assessment and model examination. For example, interactive visualization of the results of learning algorithms can help for parameter tuning. Decision trees are typical examples of models that are visualised to select the best meta-parameters (e.g. to maximise readability by non-expert end-users).

The *Integration between Visualization and Mining* opens many promising directions of research. Let us e.g. consider a high-dimensional dataset that the user wants to grasp. One could imagine that the user uses an interactive visualisation tool linked with clustering and dimension reduction algorithms. As the exploration proceeds, the user could twist the clusters by providing feedback to the clustering algorithm (e.g. "there are not enough clusters" or "those two instances should not belong to the same cluster"). In return, the dimension reduction could also adapt to better expose the clusters. Here, both the machine learning results and the visualisation change iteratively to adapt to the user.

The papers presented in this special session are representative of the three categories described above. Rayar et al. [17] present a tool for the visualisation of large image collections using a clustering algorithm, which is a good representative of the *Computationally enhanced Visualization* category. Another example is the article presented by Turkay et al. [18], which presents a tool to help social scientists build their models. Visualisations are used at each step to ensure that social scientists have a complete understanding over how their model is built.

On the other side of the spectrum, the work on interactive dimensionality reduction presented by Díaz et al. [19] fits well the *Visually enhanced Mining* category. The article presented by Barron et Whitehead [20] takes it a step further in integrating machine learning and information visualisation by providing a solution for visualising the features of unsupervised deep networks. A relatively comparable example is the article of De Bie et al. [21] which presents a framework for more meaningful data projections for high-dimensional data.

Finally, some works fall into the *Integration of Visualization and Mining* category. On that regard, the framework presented by Sacha et al. [22] is of particular interest to both communities. This conceptual framework models human interactions with machine learning components in a visual analysis process, with examples, and ends with three open research challenges at the intersection of machine learning and information visualization research.

References

[1] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[2] Anne Treisman. Preattentive processing in vision. *Computer vision, graphics, and image processing*, 31(2):156–177, 1985.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

[3] Alan Dix and Geoffrey Ellis. Starting simple: adding value to static visualisation through simple interaction. In *Proceedings of the working conference on Advanced visual interfaces*, pages 124–134. ACM, 1998.

[4] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[5] Daniel A Keim. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1–8, 2002.

[6] Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel. Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081). *Dagstuhl Reports*, 2(2):58–83, 2012.

[7] A. A Freitas. Are we really discovering interesting knowledge from data? *Expert Update*, 9(1):41–47, 2006.

[8] S. Rüping. *Learning interpretable models*. PhD thesis, Universität Dortmund, 2006.

[9] A. A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

[10] H. Allahyari and N. Lavesson. User-oriented assessment of classification model understandability. In *Proc. SCAI*, pages 11–19, Trondheim, Norway, 2011.

[11] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić. Comprehensibility of classification trees–survey design. In *Proc. IS*, pages 70–73, Ljubljana, Slovenia, 2014.

[12] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić. Comprehensibility of classification trees - survey design validation. In *Proc. ITIS*, pages 5–7, Šmarješke toplice, Slovenia, 2014.

[13] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

[14] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

[15] Daniel A Keim, Tamara Munzner, Fabrice Rossi, and Michael Verleysen. Bridging information visualization with machine learning (dagstuhl seminar 15101). *Dagstuhl Reports*, 5(3), 2015.

[16] Enrico Bertini and Denis Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Explorations Newsletter*, 11(2):9–18, 2010.

[17] Frédéric Rayar, Sabine Barrat, Fatma Bouali, and Gilles Venturini. Incremental hierarchical indexing and visualisation of large image collections. In *Proc. 24th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, 2016.

[18] Cagatay Turkay, Aidan Slingsby, Kaisa Lahtinen, Sarah Butt, and Jason Dykes. Enhancing a social science model-building workflow with interactive visualisation. In *Proc. 24th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2016.

[19] Ignacio Díaz, Abel A. Cuadrado, and Michel Verleysen. A state-space model on interactive dimensionality reduction. In *Proc. 24th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2016.

[20] Trevor Barron and Matthew Whitehead. Visualizing stacked autoencoder language learning. In *Proc. 24th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2016.

[21] Tijl De Bie, Jefrey Lijffijt, Raúl Santos-Rodríguez, and Bo Kang. Informative data projections: A framework and two examples. In *Proc. 24th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2016.

[22] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. Human-centered machine learning through interactive visualization. In *Proc. 24th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2016.