Spreadsheet interfaces for usable machine learning

Advait Sarkar
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge, United Kingdom
advait.sarkar@cl.cam.ac.uk

***Abstract*—In the 21st century, it is common for people of many professions to have interesting datasets to which machine learning models may be usefully applied. However, they are often unable to do so due to the lack of usable tools for statistical non-experts. We present a line of research into using the spreadsheet — already familiar to end-users as a paradigm for data manipulation — as a usable interface which lowers the statistical and computing knowledge barriers to building and using these models.**

I. Introduction

The ability to build and apply powerful statistical models (e.g. neural networks, linear regression, and decision trees), is currently only available to users with expertise in both programming as well as statistics and machine learning. However, these tools can be beneficial to a much broader class of end-users. Professionals in a wide variety of industries such as retail, education, transportation, sports, research, etc., often require statistical analyses and predictive modelling capabilities. These are typically performed by expert statisticians known in industry as 'data scientists.'

Unfortunately, the services of data scientists are not widely available and not always accessible to those with data in need of analysis. This creates new digital divides [1], where those with the access and ability to model data are better able to take advantage of the knowledge economy. Clearly this is an issue; the scarcity of data analytics tools aimed towards non-experts in computing and statistics excludes large groups of end-users from the ongoing analytics revolution arising from rapidly increasing compute power at rapidly decreasing costs. The claim that 'everyone should have access to data analytics,' might appear excessively idealistic, as of course, analytics is a complex and skilled activity. The production of a data scientist requires years of higher education to lay the necessary foundational knowledge of mathematics and computation, and then further years of experience in the domain of the data they analyse — no amount of clever interface design can overcome these requirements. However, we are not aiming to replace the entire statistical profession, but rather intending that through our tools, a wider range of end-users will be able to address data-related issues which they previously could not.

Our proposed solution is to leverage the expressive power of spreadsheets to bring these tools to non-experts. Since end-users are likely to be readily familiar with the manipulation of data in spreadsheets, integrating statistical modelling into the spreadsheet environment appears to be a promising avenue. We frame the problem as one of end-user programming (EUP) [2]. Much work has been done on improving the quality of EUP in spreadsheets, with mature debugging approaches such as What-You-See-Is-What-You-Test.

II. Relationship to previous work

The perspectives we have adopted in this line of research emerge from a thriving academic interest in interactive machine learning (IML), which can be viewed as an instance of EUP. IML systems have been developed for several application domains [3] such as image segmentation, music composition, image labelling, and email classification.

Surprisingly, close research attention has not yet been paid to bringing IML to the spreadsheet, which is arguably the most ubiquitous and important end-user programming environment. Nonetheless, the computing industry has taken a keen interest in augmenting spreadsheets with IML-like capabilities. Recently, the incorporation of a 'flash fill' feature into Microsoft Excel [4] has shown that programming-by-example (albeit for simple string manipulation functions only) is recognised as sufficiently useful for it to be promoted as a headline, consumer-facing feature of a mass-market spreadsheet product. Flash fill operates as a 'black box': interaction is limited to simply typing into the spreadsheet; the end-user is not exposed to the model or its parameters. The restricted synthesis vocabulary of string manipulation functions also means that flash fill's modelling capabilities are severely limited in comparison to those offered by general machine learning algorithms. The interface is therefore inapplicable for complex modelling tasks.

Applications with more explicit support for statistical modelling are available, e.g., Oracle BI [5]. Subsequent to our work with Teach and Try [6], solutions from several vendors have emerged, including Microsoft's Azure ML [7] and Google's 'smart autofill' plugin for spreadsheets [8]. However, these tools are aimed towards professionals with some formal understanding of the concepts underlying statistical modelling, and are not sensitive to the needs of non-expert users. Thus, while they are more accessible than the traditional, programmatic tools for data scientists (e.g., Weka [9] and R [10]), they still fall short of the intelligibility and usability requirements of our target population, namely, end-users who do not have formal training in statistics or computing.

III. Visual machine learning in spreadsheets

So far, we have shown that it is possible to enable non-expert end-users to confidently conduct simple machine learning-driven analyses by themselves, if they are given interfaces which take advantage of familiar spreadsheet interaction metaphors. In particular, the Teach and Try system [6] demonstrated how using spreadsheet selections to demarcate training and testing data is a quick-to-learn interface technique which causes end-users to develop an appreciation for statistical inference and its limitations.

Sarkar, A. (2015). Spreadsheet interfaces for usable machine learning. In 2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC) (pp. 283–284). IEEE.
DOI: 10.1109/VLHCC.2015.7357228

More recently, the BrainCel application [11] suggests how we might further improve user involvement in the modelling process, using visualisations of the model along with meta-information such as confidence and class representation [12]. With these model and metamodel visualisations, we enable the user to perform more nuanced activities, including judging the quality of the model, identifying good training examples, and querying why and how the model makes certain predictions. Our initial testing showed that this interface enabled end-users with no statistical training to build k-nearest neighbours (k-NN) models on various datasets, and to apply these models to make predictions for incomplete data.

IV. Proposed work

Our initial work in this space has been an encouraging indicator that interactive visual machine learning in spreadsheets is viable. However, three important issues remain to be addressed, which are described as follows:

1) *Visualising models other than k-NN*: the network visualisation in the BrainCel application is a direct representation of the k-NN model, but it is not clear how the visualisation can be adapted for other algorithms. One option is to adopt visualisations which have been created on an ad-hoc basis for other models (e.g., naïve Bayes [13]). We intend to develop visualisations which facilitate the answering of *why* and *how* questions about various other common machine learning models, such as decision forests. In addition, we also intend to explore whether it might be possible to use a visualisation developed for one model to explain the behaviour of another while still maintaining consistency in the user's mental model of the system's behaviour.
2) *Additional interaction modalities*: the action of selection to indicate training and test data works well; however, it is not yet clear whether it is optimal. There are alternative approaches available. For instance: selection of certain columns only would enable the user to restrict the feature space. Another possibility is to invert the order in which the actions are taken, so that a mode is selected before making the selection, allowing the user to 'paint' regions of the spreadsheet as training or test data. We will better study these alternative interaction modalities through a series of controlled experiments.
3) *New evaluation frameworks*: it is difficult to devise powerful, general evaluation criteria for IML systems due to the complexity of the tasks they facilitate. The 'learning barriers' framework [14], which categorises barriers users encounter when learning new programming systems, is a mature approach from EUP which can potentially be used to evaluate IML systems. However, in our studies, it has emerged that the same type of learning barrier can manifest in qualitatively different ways; some barriers are 'higher' than others. Thus it may be possible to argue between two interfaces where users encounter a similar number of barriers, that one is more desirable because the barriers encountered were 'harder' or more 'sophisticated.' We will better study this phenomenon with the aim of establishing the generality of this finding.

V. Conclusion

In this paper we have outlined the case for simple and ubiquitous tools enabling end-users to build and apply machine learning models to their own data. We propose exploiting the spreadsheet, a data manipulation environment likely to be familiar and intuitive, as an ideal interface for machine learning. Our work so far has shown that it is possible to implement such tools for end-users with no formal training in statistical modelling, machine learning, or computing, enabling them to create and confidently use sophisticated machine learning models. Finally, we have described three important research questions; answering these questions would enable us to expand our prototypes into more powerful and usable tools, and establish new paradigms for evaluating such systems.

References

[1] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.

[2] A. J. Ko, R. Abraham, L. Beckwith, A. Blackwell, M. Burnett, M. Erwig, C. Scaffidi, J. Lawrance, H. Lieberman, B. Myers *et al.*, "The state of the art in end-user software engineering," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 21, 2011.

[3] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[4] S. Gulwani, "Automating string processing in spreadsheets using input-output examples," in *ACM SIGPLAN Notices*, vol. 46, no. 1. ACM, 2011, pp. 317–330.

[5] M. M. Campos, P. J. Stengard, and B. L. Milenova, "Data-centric automated data mining," in *Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on*. IEEE Computer Society, 2005, pp. 8–15.

[6] A. Sarkar, A. F. Blackwell, M. Jamnik, and M. Spott, "Teach and try: A simple interaction technique for exploratory data modelling by end users," in *Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on*. IEEE, Jul. 2014, pp. 53–56.

[7] D. Chappell, "Introducing Azure Machine Learning: A Guide For Technical Professionals," http://download.microsoft.com/download/3/B/9/3B9FBA69-8AAD-4707-830F-6C70A545C389/Introducing_Azure_Machine_Learning.pdf, 2015, last accessed: October 26, 2015.

[8] K. Davydov and A. Rostamizadeh, "Smart Autofill - Harnessing the Predictive Power of Machine Learning in Google Sheets," http://googleresearch.blogspot.com/2014/10/smart-autofill-harnessing-predictive.html, 2014, last accessed: October 26, 2015.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[10] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org

[11] A. Sarkar, M. Jamnik, A. F. Blackwell, and M. Spott, "Interactive visual machine learning in spreadsheets," in *Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on*. IEEE, 2015, pp. 159–163.

[12] A. Sarkar, "Confidence, command, complexity: metamodels for structured interaction with machine intelligence," in *Proc. 26th Annual Conference of the Psychology of Programming Interest Group (PPIG 2015)*, Jul. 2015, pp. 23–36.

[13] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 126–137.

[14] A. J. Ko, B. Myers, H. H. Aung *et al.*, "Six learning barriers in end-user programming systems," in *Visual Languages and Human Centric Computing, 2004 IEEE Symposium on*. IEEE, 2004, pp. 199–206.