Vassilis Kostakos, University of Melbourne
Mirco Musolesi, University College London and the Alan Turing Institute

Avoiding Pitfalls When Using Machine Learning in HCI Studies

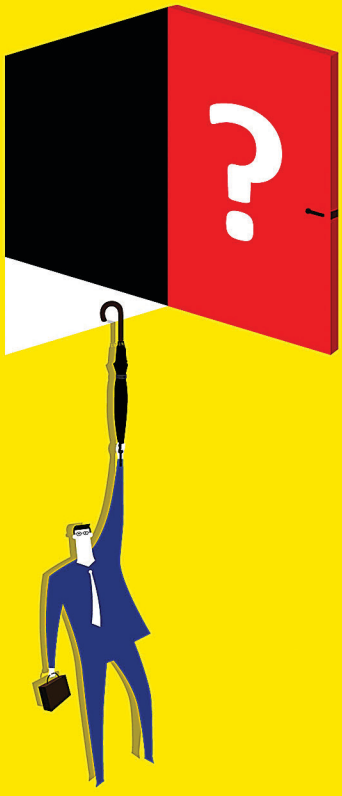
Insights

- We highlight some of the pitfalls that HCI researchers should avoid while using ML techniques.
- Prediction accuracy cannot be used as a substitute for classic hypothesis testing and correlation/causation analysis.
- In addition to accuracy, researchers should also report baseline performance.

Machine learning (ML) has come of age and has revolutionized several fields in computing and beyond, including human-computer interaction (HCI). Human-subject studies have been adopting ML techniques for more than a decade, for example for activity recognition and wearable computing. There now also exists a plethora of application domains in which ML approaches are enriching interactive computing research. Here we wish to highlight some of the pitfalls that HCI researchers should avoid while using ML techniques in their research.

A popular use of ML techniques in HCI is to model human behavior. One potential risk is that ML techniques are sometimes used inappropriately to draw conclusions, possibly strong, about human behavior, instead of using more classical statistical methods. It is worth noting here that some ML techniques are actually based on classical statistical methods such as regression or curve fitting. However, some classification methods, such as neural-networks-based approaches, are much more difficult to interpret given the complexity and dimensionality of the underlying mathematical models inferred from the data.

Another popular use of ML in HCI is to develop novel user-interface



techniques, such as to react to user input (e.g., gesture recognition), optimize system resources (e.g., smartphone battery conservation [1]), or provide intelligent mobile notifications [2]. The prediction of future users' activities and interactions is another emerging area of interest: The aim is to develop full-fledged anticipatory computing systems [3]. Indeed, a rigorous performance evaluation of these systems is fundamental in order to evaluate their effectiveness and efficiency.

Specifically, the definition of the training set needs to be considered in detail when ML techniques are used in HCI. Interactive systems are usually evaluated with a training set obtained from a certain population of users. When evaluating the system, authors should report both: results using training data only from the same individual (personalized model) and results using data from the entire population (generic model). This is necessary for systems where no data exists for first-time users and, therefore, classifiers have to be bootstrapped with data from other users. It might also be helpful to show variations in the performance for the entire population in order to understand if, for example, there are classes of users that are easier to model and predict. Sometimes the application of clustering techniques might be necessary to identify users who share the same characteristics.

ML IS NO SILVER BULLET FOR HCI RESEARCH

Classification accuracy is not hypothesis testing. It is important to underscore that ML prediction accuracy cannot be used as a substitute for classical hypothesis testing and correlation/causation analysis, especially when deriving conclusions about characteristics of human behavior. Let us consider, for example, an application for classifying the mood level of

IMAGE BY PAN JJ / SHUTTERSTOCK.COM

a person from certain behavioral characteristics. In analyzing their results, researchers have to be very careful in interpreting how these behavioral characteristics are linked to the actual emotional states of users.

Some ML methods provide insights about the interpretation of the phenomena under observation. For example, in the case of descriptive methods (such as the classic association-rule algorithm [4]), it is possible to derive potential interpretations of the observed data. However, this is not the case for other state-of-the-art algorithms, such as deep-learning techniques [5]. Although the interpretation of deep-learning-algorithm output is an area of intense research, the currently available tools provide limited information about the inner workings of the models. At the same time, it is interesting that the analysis of the output from the intermediate steps of these multi-layer architectures might provide some suggestions for isolating interesting behavioral patterns in the data.

We argue that researchers should consider using hypothesis-testing approaches in these cases to generate new knowledge about the world. These approaches may seem outdated, and in fact may be less accurate at describing the observed phenomena. However, they do offer researchers complete control over their inner workings, and therefore provide a form of language that researchers can use to construct and test hypotheses, and therefore interpret phenomena. We believe that these are essential as preliminary tests before adopting ML techniques for estimation and prediction.

So far, we have implicitly assumed that the ML algorithms taken into consideration were ones that involved supervised learning, meaning that the scientist can provide labeled data for training. We should be even more careful in the interpretation of the results from unsupervised techniques, where scientists do not have labeled data to begin with, and therefore the interpretation of the results cannot be directly guided by existing examples. One should consider, for example, the stability of the results with different parameters (e.g., in the case of topic models).

We would also like to stress the importance of visualization in interpreting behavioral data. Visualization techniques can be extremely important not only for understanding raw data, but also for interpreting (fitted) models derived from the application of ML techniques, for example through projections of highly dimensional models.

Causality versus correlation. Another important aspect to consider is the problem of correlation versus causation. Most of the results of ML algorithms provide insights into association relationships and not causality relationships. Consequently, researchers should be extremely careful in extrapolating conclusions from results that might be the effect of correlation and not causation. This is not a new problem, but it is exacerbated by the fact that nowadays many studies are based on data collected through crowdsourcing, third-party APIs (such as the Twitter API), and mobile apps distributed in Web stores and open to the public. It is also worth noting that causality is a very active area in the ML community at the moment. We expect that many tools will be made available to practitioners in the years to come.

Controlled versus non-controlled experiments. Different techniques should be used in controlled versus non-controlled experiments. Indeed, it is important to be very careful in drawing conclusions from experiments that rely on non-controlled designs, for example systems for positive behavioral intervention. Having said that, there are well-established methods proposed by the ML and statistics communities for dealing with unbalanced populations. In other words, it is possible to analyze non-controlled experiments, but researchers have to be very careful in the analysis of their results and in drawing appropriate conclusions. In non-controlled experiments, causality analysis is very difficult but not impossible, for example, if quasi-experimental approaches are applied [7]. Indeed, it is interesting to note that in many application scenarios, quite often it is simply impossible to build control groups when data is crowdsourced or collected through mobile applications distributed on Google Play or the Apple App Store. This is an area of great interest not only for the ML/statistics community but also in other disciplines, for example health studies, epidemiology, and geo-demographics.

Although the interpretation of deep-learning-algorithm output is an area of intense research, the currently available tools provide limited information about the inner workings of the models.

HOW GOOD IS GOOD ENOUGH? AND WHAT DO WE MEAN BY *GOOD*?

There seems to exist an unwritten convention that classifiers with accuracy above 80 percent are "good enough" and therefore publishable. Yet there is little consistency in how HCI researchers interpret classifier accuracy, and in fact how they *report* classifier accuracy. We argue that in addition to accuracy, researchers should also report baseline performance.

Consider a system that attempts to infer the *gender* of a user by analyzing their mobility habits. In this case, there are two possible outcomes (male and female), and therefore we can assume that a baseline performance is 50 percent (e.g., reflecting the toss of a random coin). Classifier performance is judged against this baseline, and therefore a classifier that performs at 85 percent accuracy improves the baseline by a factor of 0.7. Alternatively, a gesture-recognition system that differentiates between 15 different gestures has a baseline performance of 1/15, or 6.6 percent. If such a system achieves accuracy of 85 percent, then it is improving the baseline by a factor of 11.9. Hence, interpreting the accuracy of a classifier needs to be set against a (random) baseline. And, actually, we argue that often accuracy results around 30 to 40 percent might already

be considered excellent in the case of the difficult classification problems described earlier. For this reason, it is fundamentally important to discuss performance always in relation to the complexity of the ML task under consideration (and, indeed, of the state of the art in the field!).

Furthermore, especially in behavioral studies, it is important to note that the baseline is a function of both the possible outcomes and the relative likelihood of each. For instance, consider a system that monitors all the sensors on the smartphone and attempts to predict whether a user is going to answer their phone if someone calls. Even if we assume only two possible outcomes (answer, no answer), the baseline is not necessarily 50 percent. This is because we may observe that, overall, users almost always answer their phone when it rings. If, say, we observe that 90 percent of the time the user answers the phone, then this also acts as our baseline: If we construct a classifier that constantly predicts that the phone will be answered, its accuracy will be 90 percent. In this case, if a study reports their classifier performing at 85 percent, it is actually performing *worse* than the baseline. The actual baseline should then be not a purely random case, but rather a *frequency-based* classifier.

Finally, it is worth noting that accuracy is not sufficient to evaluate ML classification algorithms. For example, the existence of false positives is another very important aspect that is often not sufficiently considered in the evaluation of studies that rely on ML techniques. A false positive is the result of a test that indicates a certain finding or condition exists when it actually does not. An example is the case of a classifier that reports that a user can be interrupted at a certain point in time, when in fact the ground-truth data demonstrates this is not the case. A true positive instead is a result of a test that indicates the condition is actually verified. Indeed, it is necessary to report indicators expressing the sensitivity (i.e., the proportion of positives that are classified as positives) and specificity (i.e., the proportion of negatives that are classified as negatives) of the results. In the case of binary classifiers, for example, standard evaluation techniques include the use of the Receiving Operating Characteristic curve (ROC curve) and the Area Under the (ROC) Curve (usually abbreviated as AUC). ROC curves are used to evaluate the specificity and sensitivity of a classifier considering different threshold settings of the classifiers. The discussion of these techniques is beyond the scope of this article; for an excellent step-by-step discussion of these and other evaluation strategies for ML techniques, we refer the reader to [6].

CONCLUSION AND OUTLOOK

We believe that ML offers immense opportunities to HCI researchers. However, just as in performing statistical modeling, we should constantly remind ourselves of caveats in the analysis ("correlation does not mean causality"). Today too we must embrace ML approaches while having a keen understanding of their current limitations and prospects for improvement in the near future.

It is also worth noting that nowadays a large number of tools and libraries for ML are available as stand-alone tools (e.g., Weka), R libraries (e.g., randomforest), or Python libraries (e.g., scikit-learn). We believe that, even if it is not important for HCI researchers to understand how the tools work, it is essential to have a general knowledge of the underlying algorithms and key parameters—the "knobs" of the algorithms—both for improving their performance and for understanding the data. For these reasons, we argue that a solid background in the basics of ML is necessary before adopting these tools in our research work and practice. Related to this, it is interesting to note that various universities have introduced (or will introduce) an introduction to ML concepts and techniques as part of advanced courses in HCI and/or ubiquitous computing.

Finally, we also believe that qualitative methods must play a fundamental role in interpreting quantitative data obtained by means of quantitative methods such as the application of ML techniques. A mixed-methods approach is usually the most promising when interpreting human behavioral data, which is inherently complex, noisy, and incomplete. Moreover, often ML techniques are applied to subsets of the data and, therefore, the resulting models capture only a limited part of the phenomena under observation.

In this article we have attempted to highlight some issues that are becoming increasingly important within HCI research and offer some material as a basis for starting a discussion in the community around these themes. We have emphasized the importance of understanding the subtleties in using these techniques and tools, while keeping in mind the exceptional opportunities deriving from their adoption in our research work.

Endnotes

1. Kostakos, V., Ferreira, D., Goncalves, J., and Hosio, S. Modelling smartphone usage: A Markov state transition model. *Proc. of the 2016 International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, 2016.
2. Mehrotra, A., Musolesi, M., Hendley, R., and Pejovic, V. Designing content-driven intelligent notification mechanism for mobile applications. *Proc. of the 2016 International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, 2015.
3. Pejovic, V. and Musolesi, M. Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Computing Surveys 47*, 3 (Apr. 2015).
4. Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*. ACM. New York, 1993.
5. Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2017.
6. Flach, P. *Machine Learning. The Art and Science of Algorithms that Make Sense of Data*. Cambridge Univ. Press, 2012.
7. Tsapeli, F. and Musolesi, M. Investigating causality in human behaviour from smartphone sensor data: A quasi-experimental approach. *EPJ Data Science 4*, 1 (2015).

Vassilis Kostakos is a professor of computer engineering at the University of Melbourne. His research interests include ubiquitous computing, human-computer interaction, and social computing. He has a Ph.D. in computer science from the University of Bath.
→ vassilis.kostakos@unimelb.edu.au

Mirco Musolesi is a reader in data science at UCL and Faculty Fellow at the Alan Turing Institute. His research interests lie at the intersection of ubiquitous computing, mobile sensing, large-scale data mining, and network science. He has a Ph.D. in computer science from University College London.
→ m.musolesi@ucl.ac.uk

DOI: 10.1145/3085556