Avoiding Pitfalls in Using Machine Learning in HCI Studies

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

Machine Learning (ML) has come of age and has revolutionized several fields in computing and beyond, including Human Computer Interaction. Historically, human subjects studies have adopted ML techniques for more than a decade, for example for activity recognition and wearable computing. However, there now exists a plethora of application domains where 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.

Machine Learning in HCI

Human Behaviour Analysis

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

User Interface Techniques

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 [4]) or provide intelligent mobile notifications [5]. The prediction of future user's activities and interactions is another emerging area of interest: the aim is to develop full-fledged anticipatory computing systems [6]. Indeed, a rigorous performance evaluation of these systems is fundamental in order to evaluate their effective and efficient.

Specifically, the definition of the training set is an aspect that 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 that from the entire population (generic model). This is necessary for systems where no data exist for first-time users and, therefore, classifiers have to be bootstrapped with data from other users. It might also be helpful to show distributions of the sensitivity/specificity 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 in order to identify users that share the same characteristics.

ML is no Silver Bullet for HCI Research

Classification Accuracy is not Hypothesis Testing

It is important to underline that ML prediction accuracy cannot be used as substitute for classic 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 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 actual interpretation of the phenomena under observation. For example, in the case of descriptive methods (such as the classic "association rule" algorithm [1]), 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 [3]. Although the interpretation of deep learning algorithm output is an area of intense research, the current available tools provide limited information about the "inner workings" of the models. At the same time, it is interesting to note 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 of the world. 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 supervised learning ones, 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 labelled 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 (for instance in the case of topic models).

Finally, 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 by the application of ML techniques, for example through projections of highly-dimensional ones.

Causality vs Correlation

Another important aspect to consider is the problem of correlation vs causation. Most of the results of ML algorithms provide insights about association relationships and not on 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 and we expect that many tools will be made available to practitioners in the years to come.

Controlled vs Non-Controlled experiments

Different techniques should be used in controlled vs. 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 [5]. Indeed, it is interesting to note that in many application scenarios quite often it is simply impossible to build control groups when data are crowdsourced or they are collected through mobile applications distributed on Google Play or 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, geo-demographics just to name a few.

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% are "good enough" and therefore publishable. Yet, there is very 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 2 possible outcomes (male, female), and therefore we can assume that a "baseline" performance is 50% (e.g. reflecting the toss of a "random coin"). Classifier performance is judged against this baseline, and therefore a classifier that performs at 85% 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 = 6.6%. If such a system achieves accuracy of 85%, 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 30/40% might already be considered as excellent in case of difficult classification problems as that described above. 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 not only a function of the possible outcomes, but also the relative likelihood of each. For instance, consider a system that monitors all 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 2 possible outcomes (answer, no answer) the baseline is not necessarily 50%. This is because we may observe that, overall, users almost always answer their phone when it rings. If, say, we observe that 90% 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%. In this case, if a study reports their classifier performing at 85%, it is actually performing *worse* than the baseline. The actual baseline should then not be a purely random case, but a *frequency-based* classifier.

Finally, it is worth noting that accuracy is not sufficient to evaluate ML classification algorithms. In fact, 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 that a certain finding or condition exists when it actually does not. An example is the case of a classifier that reports that a user is focusing on a certain point in mobile webpage when she is not looking at the ground-truth data. A true positive instead is a result of a test that indicates that 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 case of binary classifiers, for example, standard evaluation techniques include the use of Receiving Operating Characteristic curve (ROC curve) and the Area Under the (ROC) Curve (usually abbreviated with 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 outside the scope of this review; for an excellent step-by-step discussion of these and other evaluation strategies ML techniques, we refer the reader to [2].

Conclusions and Outlook

We believe that ML offers immense opportunities to HCI researchers. However, just as in performing statistical modelling, 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, i.e., the knobs of the algorithms, both for improving their performance, but also 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 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 are inherently complex, noisy and incomplete. Moreover, often ML techniques are applied to subsets of the data and, therefore, the resulting models only capture 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 underlined the importance of understanding the subtleties in using these techniques and tools, while, at the same time, keeping in mind the exceptional opportunities deriving from their adoption in our research work.

References

[1] Rakesh Agrawal, Tomasz Imielinski and Arun Swami. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of data (SIGMOD'93). ACM.

[2] Peter Flach. Machine Learning. The Art and Science of Algorithms that Make Sense of Data. 2012. Cambridge University Press.

[3] Ian Goodfellow, Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press. 2017.

[4] Vassilis Kostakos, Denzil Ferreira, Jorge Goncalves, Simo Hosio. Modelling Smartphone Usage: A Markov State Transition Model. Proceedings of the 2016 International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16). 2016. ACM

[5] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley and Veljko Pejovic. Designing Content-driven Intelligent Notification Mechanism for Mobile Applications. Proceedings of the 2016 International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15). 2015. ACM.

[6] Veljko Pejovic and Mirco Musolesi. Anticipatory Mobile Computing: A Survey of the State of the Art and Research Challenges. In ACM Computing Surveys. ACM. Volume 47. Issue 3. April 2015.

[7] Fani Tsapeli and Mirco Musolesi. Investigating causality in human behaviour from smartphone sensor data: a quasi-experimental approach. EPJ Data Science. Volume 4. Number 1. 2015. Springer.