Interactive Data Mining Considered Harmful* (If Done Wrong)

Pauli Miettinen
Max-Planck-Institut für Informatik
Saarbrücken, Germany
pauli.miettinen@mpi-inf.mpg.de

ABSTRACT

Interactive data mining can be a powerful tool for data analysis. But in this short opinion piece I argue that this power comes with new pitfalls that can undermine the value of interactive mining, if not properly addressed. Most notably, there is a serious risk that the user of powerful interactive data mining tools will only find the results she was expecting. The purpose of this piece is to raise awareness of this potential issue, stimulate discussion on it, and hopefully give rise to new research directions in addressing it.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Interactive data analysis; statistical testing; white paper

1. INTRODUCTION

Traditionally, the KDD process was presented as a waterfall, going from pre-processing to data mining to post-processing (solid lines in Figure 1). This—of course—has never been true, and more modern models of data mining, such as Shearer's CRISP-DM model [12], reflect that. Data analysis is an iterative process: the user prepares the data, selects analysis methods and their parameters, runs the methods, studies the outcome, and returns to any of the earlier steps, possibly preparing the data differently, or using different analysis method or different parameters (dashed lines in Figure 1).

But this iterative process is arduous and each step that needs to be repeated can take a significant amount of time. To help with this is what the *interactive* data mining is for: to allow the user to pinpoint the analysis method to the interesting results without the time-consuming iteration. Done well, interactive data mining methods can be extremely powerful, giving the user unprecedented machinery to better understand her data. But with great power comes great responsibility, as the saying goes. By allowing the user to control the data mining process in (near) real time, interactive data mining systems posses the risk of undermining the very promise of data mining: discovering new and unexpected knowledge.

*With apologies to Edsger W. Dijkstra

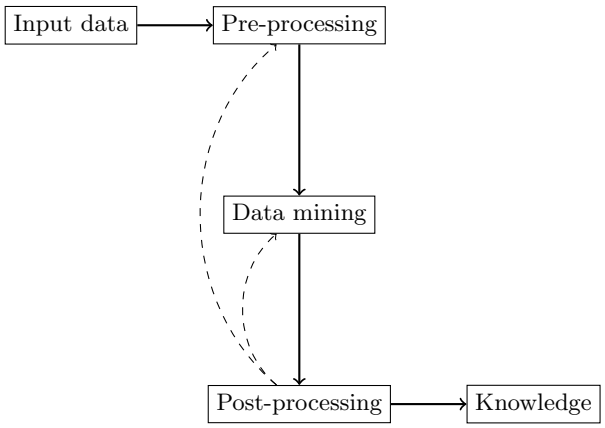

Figure 1: The iterative KDD process

2. THE PROBLEM

The goal of data mining, in the words of one textbook, is

> [T]o find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. [4]

Data mining community has always been good at inventing novel ways to mine the data, but has perhaps struggled more with the understandability and usefulness parts. It is these two areas that interactive data mining tries to improve by letting the user to tell the algorithm, during the mining process, what she finds useful and understandable. But doing so, it threatens a very important aspect of data mining mentioned in the above quote: the results should be *unsuspected*.

The user using the interactive data mining method is (hopefully) familiar with the data and what it represents. Consequently, the user has some prior ideas what the potential results could be, and what kind of a result would be a useful result in this domain. But these prior ideas might—indeed, I argue they will—make the user steer the algorithm towards the kind of results that she a priori considered useful and interesting, and never find the kind of results she did not expect to find. This can make the interactive data mining, intended to be exploratory by nature, a confirmatory data analysis technique—and not necessarily very good method at that, even.

To give a more concrete example, consider an interactive data mining algorithm that presents the user with partial results in an anytime fashion and lets her to guide the search with feedback such as "more like this" or "less like this."

Contemporary interactive data mining methods might not quite achieve this level of interaction yet, but it is clear that it would be desirable if they would. It should be obvious, however, how the user can, possibly unintentionally, use this feedback mechanism in such a way that the algorithm only returns results that she was expecting.

3. IS IT A (NEW) PROBLEM AT ALL?

Is this a real problem? Is it not a far-fetched idea that the user would have on her mind the exact results the mining algorithm will find? It indeed is, but it is important to note that this problem appears as soon as the user has even a vague a priori idea on what would be a useful result from the algorithm. And for a user with only a faint idea on what could be useful, what is the purpose of interactive data mining, what is its added value? The potential lose of surprising results is the price to pay for the power of interaction, the Jekyll and Hyde of interactive data mining.

But has this problem not been part of data mining all the time? As already discussed, the process of knowledge discovery is iterative and the user can repeat the steps trying to extract more understandable and useful results, potentially removing the more surprising results while doing so. But interactive data mining tools can emphasize this problem significantly by giving the user a faster access to the mining process; indeed, interactive, rather than iterative, access. Again, the problem lies in the heart of interactive data mining: the power that interactive data mining gives to the user over the iterative data mining is exactly the same power that lets the user to only find the unsurprising results.

The users, one could argue, would not intentionally avoid the unsuspected results. But oftentimes, it is hard to appreciate such results in the first glance. The results, being unsuspected, might look like noise or random occurrences as they do not fit into our thinking of the data. They might require us to update our understanding of the data, possibly running more experiments, before we can appreciate them, all of which makes the process significantly less interactive. Yet, it is precisely the change in understanding the data these results require that makes them so valuable for the mining process.

A related problem in statistics and machine learning is that of over-fitting. By steering the data mining process away from unsuspected results, the user is effectively over-fitting the results into her prior assumptions. But this kind of over-fitting is much harder to address than the more common one. The final arbitrator for the quality of a machine-learning algorithm is its predictive power. But data mining is descriptive, rather than predictive, and in many cases, there is no clear prediction stemming from the results. There is no objective quality measure, either, as here the user is the arbitrator of the quality.

4. POSSIBLE SOLUTIONS

Arguably the simplest solution is user education. The power to interact with the algorithm is vested in the user, and she should be taught how to use this power. Unfortunately, education alone cannot solve all the problems. The risk of missing important but unsuspected results exists whenever the user is allowed to interact with the algorithm, any education notwithstanding, and if this power is removed from the user, there is not much interactive data mining left.

Another simple approach is to restrict the power of the interaction, keeping the situation closer to status quo. It should go without saying that this approach is sub-optimal.

The potential for data mining algorithms, and their users alike, to concentrate on "wrong" results has existed all the time. Significant amount of data mining research is devoted to testing whether a specific result is significant with respect to some null hypothesis (e.g. [2, 3, 8, 9, 11]) or even with respect to user's prior knowledge (e.g. [1, 5, 10]), to say nothing about the vast body of statistical literature on measuring the statistical significance. In principle, the approach these papers take can be used to steer the user and the algorithm away from expected results: encode the users prior knowledge in the null hypothesis and discard results that are not significant under this null hypothesis (and interactively update the null hypothesis when new results are obtained).

While the general approach of using significance testing is very appealing, it is not clear at all whether it can be used to actually alleviate the problem in the interactive setting. First, the significance testing must be instantaneous—or at least fast enough to be used interactively. Some methods, for example the maximum entropy methods, should be able to pass this hurdle, while others, such as permutation test style swap randomization, most probably will not. Second, the user should be able to communicate her a priori assumptions to the method so that they can be build in to the null hypothesis. Given that even a vague prior belief can have a negative effect, this might be too tall an order. It could be circumvented to some extend by simply relying on the interactive nature of the algorithm: updating the null hypothesis based on user's interaction with the algorithm and her reactions to the new results could reveal enough of her latent a priori assumptions for the method to work.

The biggest hurdle for this method, however, is in its very nature: significance testing is designed to spot insignificant results, but it does not, per se, help at finding new significant results. For example Mampaey et al.'s method [10] rely on clever algorithms to actually find the patterns. Should such algorithm be endowed with "more like this/less like this" kind of functionality, there would still be nothing stopping the user from steering the algorithm away from unsuspected results. It could well happen that the user would find almost nothing of significance: her own actions would guide the algorithm away from the unsuspected results, while the significance testing would deem almost all of the remaining results redundant or insignificant with respect to the prior knowledge.

In fact, it might well be that there is no (computationally efficient) solution to the problem, at least not unless we place strong assumptions on the users' behavior. In the statistical query model of Kearns [7], the user asks questions about the expected value of a predicate over some (finite) distribution. The algorithm, called oracle, does not know the distribution, but has access to a sample of size n from it. The algorithm's task is to give valid answers, that is, answers that do not deviate too much from the true expectation, based only on the sample. In their recent paper, Hardt and Ullman [6] showed that there is no computationally efficient algorithm that can give valid answers to $n^{3+o(1)}$ *adaptive* statistical queries assuming one-way functions exist[1].

[1]A *one-way function* is, informally, a function which is easy to compute for any input, but hard to invert given an image of a random input. Their existence is a standard assumption in much of modern cryptography.

The crux in Hardt and Ullman's result is the adaptivity, as giving valid answers to even exponential number of non-adaptive statistical queries is easy. We can interpret the result in two ways: On one hand, it at least shows that adaptive queries are significantly harder to answer correctly than non-adaptive ones. On the other hand, we can interpret the result to tell even more about the computational limitations of interactive (and iterative, for that matter) data analysis systems: that it is impossible to prove that our results are even correct, to say nothing of surprising, assuming that the user can ask sufficiently many adaptive questions.

5. TESTS

The final, and perhaps the most important, piece on addressing the problem is testing it. Without testing, we do not know if the problem even exists, nor can we assess the effects of proposed solutions. Developing tests to measure if the interaction makes the users to miss unexpected results is, unfortunately, not easy. It does not seem likely that it could be tested without involving humans to act as users. A potential test could have two groups of users, a test group using the interactive algorithm, and a control group using non-interactive algorithm. Their findings would then be evaluated to measure whether the test group missed results the control group found, or vice versa. But even this seemingly simple test setup requires many design decisions to be made—where are the test subjects found, what are the group sizes, how can it be ensured that the test is fair, and how are the results interpreted—and traditionally data miners have not been the ones with best knowledge about and keenest interest on human experiments. Luckily, this is a problem that should be very easy to solve by collaborating with experts.

6. CONCLUDING REMARKS AND CALL FOR ACTIONS

Interactive data mining is a powerful form of data analysis with the potential of becoming the standard format of data mining. But it comes with new pitfalls that need to be taken into account when new interactive data mining methods are developed and analyzed, lest the results become void of unexpectedness. The community should, therefore, start addressing the problem of finding only expected results: we need methods to test the seriousness of the problem and the effects of the attempts to alleviate it; we need general frameworks to help avoiding the problem; and we need interactive algorithms that try to steer the user away from discovering only the expected results. But above all, we need to realize that this is a potential problem and start thinking about it.

7. REFERENCES

[1] T. De Bie. Maximum entropy models and subjective interestingness: An application to tiles in binary databases. *Data Min. Knowl. Discov.*, 23(3):407–446, 2011.

[2] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Trans. Knowl. Discov. Data*, 1(3), 2007.

[3] W. Hämäläinen. *Efficient search for statistically significant dependency rules in binary data*. PhD thesis, University of Helsinki, Oct. 2010.

[4] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, Massachusetts, 2001.

[5] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila. Tell me something I don't know: Randomization strategies for iterative data mining. In *KDD '09*, pages 379–388, June 2009.

[6] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS '14*, 2014. To appear.

[7] M. Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, Nov. 1998.

[8] A. Kirsch, M. Mitzenmacher, A. Pietracaprina, G. Pucci, E. Upfal, and F. Vandin. An efficient rigorous approach for identifying statistically significant frequent itemsets. *J. ACM*, 59(3), June 2012.

[9] K.-N. Kontonasios, J. Vreeken, and T. De Bie. Maximum entropy modelling for assessing results on real-valued data. In *ICDM '11*, pages 350–359, 2011.

[10] M. Mampaey, N. Tatti, and J. Vreeken. Tell me what i need to know: Succinctly summarizing data with itemsets. In *KDD '11*, pages 573–581, Aug. 2011.

[11] M. Ojala. Assessing data mining results on matrices with randomization. In *ICDM '10*, pages 959–964, 2010.

[12] C. Shearer. The CRISP-DM model: The new blueprint for data mining. *J. Data Warehous.*, 5(4), 2000.