From Patterns to Discoveries

Michael R. Berthold

1 The Past

Over the past two and half decades or so, a typical data analyst, such as myself—even though I really did not even realize at start that I was doing data analysis—started off by learning descriptive statistics and getting truly excited about the promises of machine learning. The ability of those methods to match powerful models to—today ridiculous but back then huge amounts of data opened up endless opportunities. I dove into neural networks and other similar approaches, together with their algorithms to match complex, distributed models to large data sets [1].

From this, the step to trying to actually understand those models was a natural desire. Being able to predict sufficiently ahead of time that quality in a production line will be dropping is one aspect, but being able to explain why this is the case allows to actually fix the problem at the root. So in the 1990s, I started looking into rule models and decision trees to extract interpretable patterns from large data sets. Extensions of these methods based on imprecise logic allowed to inject (inherently imprecise) expert knowledge and extract more general but still performant interpretable models [2]. This shift made it possible to find interpretable models to somewhat larger data sets, but we quickly ran into walls nevertheless.

Due to one of those strange shifts in life, I started working on data sets from the life science industries—the company hiring me to head their data analysis think tank realized that there was a wealth of powerful methods in the data mining community with potential for their own applications and was looking for fresh insights—or as their CEO put it: someone unbiased by any actual knowledge about the underlying applications and previous work in the field. In me they found

M.R. Berthold (✉)
Nycomed-Chair for Bioinformatics and Information Mining, Department of Computer and Information Science, Graduate School on Chemical Biology (KoRS-CB), University of Konstanz, Konstanz, Germany
e-mail: Michael.Berthold@Uni-Konstanz.DE

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_4,

a perfect candidate; I could barely spell "biology." When my focus shifted to life science data analysis, it became apparent that researchers in this area struggled with much more pronounced problems: they often did not even know what questions to ask! It took me a few months to abandon the holy grail of data mining: performance. In the life science areas, nobody cared about that famous last percentage point of model accuracy—the experts in the field would not trust our models anyway. They wanted explanations. And much more importantly, they wanted to see something new, something interesting, and something that would trigger a truly new discovery! During one of the long discussions we had with our users, trying to find out what kind of patterns they were looking for, a chemist finally got really nervous and exclaimed:

> I don't know what I am looking for, but I know it when I see it!

Since then, this phrase has driven much of the research in my group; see, for example, [3–6].

2 Types of Miners

It is interesting to note that the different types of data analysis described informally in the section before can be grouped into three main categories and nicely matched to different phases of scientific research.[1] Figure 1 illustrates this analogy.

2.1 Parameterization

This phase of data mining concerns essentially the fine tuning of an answer we already know exists, but we are lacking a few parameter values to really understand the underlying system—statistical data analysis represents the core technology on the data mining end. This type of analysis corresponds to the third phase in scientific research: Formalization. The system is well understood and the remaining questions rotate around fitting parameters. Theory formation and systematic experimentation go hand in hand.

2.2 Pattern Detection

The classical data mining area focuses on finding patterns. The nature of the patterns (e.g., association rules) is clear, but we do not know yet where those

[1] Personal communication with Christian Borgelt who cited (from memory) a publication that we were unable to find. Please contact the author if you know the reference.

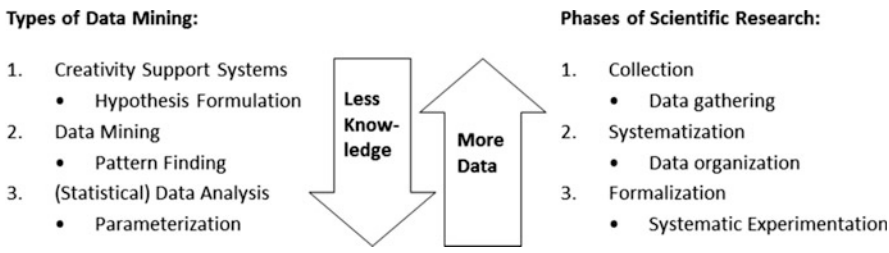

Fig. 1 The three phases of scientific research and the matching types of data mining

patterns occur. So we need highly efficient algorithms to dig through vast amounts of data to find the (often only few) local instantiations. The scientific research equivalent is conceptualization or systematization: we have an understanding of major parts of the system but lack some details. In a way, this phase relates nicely to question answering systems: can you find me more instances of this pattern? Can you determine the subset of features needed to explain this outcome? Finding a predictor is really a subdiscipline here: we may not care about the interpretability of the result, but we know (or at least we believe we do) which features may have an influence on the target and try to find some black box mimicking the underlying system's behavior.

2.3 Hypothesis Generation

The third phase of data mining activities corresponds to the early discovery phase of a new scientific area. Nothing really is known about the overall structure of the underlying system, and the goal is essentially to collect evidence wherever possible that can be used to poke holes into the fog and gradually build up a more complete understanding. This relates to the setup sketched above: the user cannot even really specify the question. Instead, we need to support users to form questions in the first place! This phase is often characterized by the ability to generate huge amounts of data and not really knowing what to do with it. The scientific research phase "Collection" stresses this aspect: we are collecting bits and pieces of information but do not (yet) know how those pieces fit into the overall puzzle. The goal is help users form new hypotheses. Quite a few new research areas arose around this topic in various fields: visual analytics in the visualization community, active learning in the machine learning sphere, and even data mining have a spin off: explorative data mining. But the overarching goal is hardly ever stressed explicitly: the aim is to build creativity support systems that help users to generate new, exciting questions.

One big problem when moving into this last phase becomes validation. Methods for parameter fitting and pattern mining can rather objectively be evaluated on standard benchmarks and using well-founded metrics. Obviously, there is a bit of a

risk of overfitting[2] but, with a bit of care, strong conclusions can be drawn from a well-done experimental analysis.

However, for the third type "hypothesis generation," this is far more complex. Finding a new hypothesis is not something that can be recreated in an isolated test bed. Measuring what is interesting is a difficult problem altogether. So the danger is that we will see—again—lots of papers presenting new approaches consisting of small modifications of existing methods and present their proclaimed superiority on one or two well-chosen examples. Better would be reports of successes on real life use cases, but are those really broad evaluation and can those really be used for comparisons? Fair validations of such systems will be an interesting challenge.

3 Tools

It is interesting to look at the evolution of development of tools for data processing, analysis, or mining over the past 20 years as well—they tend to follow the phases outlined above but typically lack behind a number of years. Maybe we can learn something from this trend for future developments.

Initially, tools were essentially table based, following the usual spreadsheet setup of VisiCalc or later mainly Excel. These tools allow the users to do increasingly complex operations but restrict those operations essentially to a single table. A parallel development was highly sophisticated statistical programming languages such as S and the open source version R. These languages allow highly skilled experts to run pretty much any analysis anyone can think of on a variety of data sets. However, the resulting code is far from being maintainable, and usually, research departments have a few experts that the analysis hinges upon. In a way, this development matched the first phase outlined above—at more or less advanced levels, of course.

A variety of other tools showed up briefly (such as SNNS, the Stuttgart Neural Network Simulator), but none really stuck around for long. One very notable exception is Weka, initially a machine learning toolkit which, when data mining research gained momentum, also became enormously popular among data miners. Weka, similar to R, managed to create a community of researchers adding new components to the platform and quickly became a reference platform among machine learning and data mining researchers. Weka still requires users to be pretty firm in the use of sophisticated algorithms and somehow exemplifies phase 2: if you are in search for a state-of-the-art data mining algorithm, you can likely find it in either Weka or—even more likely—also R. Of course, in parallel to those open source developments, also statistical analysis firms such as SPSS and SAS started to develop or buy tools similar to those resulting in tools such as Enterprise Miner and

[2] As Pat Langley once put it: "The entire Machine Learning community is kind of overfitting on the UCI Benchmark collection."

Clementine. In contrast to R and Weka, however, they did not foster community integrations and are not as heavily used by data mining researchers.

For phase 3, we are looking at a different picture: we need to enable users from other application areas to use our research prototypes instead of the usual benchmark evaluation. We can neither expect them to be aware of the latest bleeding edge development in our field nor can we expect them to make use of every little parameter twist in existing algorithms. They want to use tools, play with their data, and, by exploration, come up with new ideas and hypotheses. Those tools need to be interactive to allow exploration and—most importantly—intuitive to use also for expert as well as novice users!

This results in a dramatic shift in requirements for data mining researchers trying to trigger progress in this area: supporting complex, explorative analyses will require—even for research prototypes—fairly stable, professional grade tools. Otherwise, researchers in the application areas will simply not be able to use those in their live working environments and, in return, researchers in the data mining area will not be able to seriously evaluate their methods. This, of course, poses an entire set of new challenges for researches in the data mining field: in addition to publishing algorithms, we will now need to accompany these publications with the deployment of the new methods in environments that can be used in a professional context.

Obviously, we cannot require professional tool development from each and every research group. I personally do not see a way around highly modular frameworks that can be used by the broader research community. At the University of Konstanz, we invested over 2 years of time building up such a framework, guided by three main goals and resulting in KNIME, the Konstanz Information Miner [6][3]:

- Professional grade: from day one experienced, software engineers were part of the KNIME core team.
- Modularity: an integration platform can only be as good as its weakest piece. Therefore, KNIME essentially sandboxes all of the individual modules sanity checking as much of their operation and subsequent output as possible. This way, we can quickly determine why a workflow failed and isolate the misbehaving module.
- Community involvement: in order to allow others to (a) benefit from the work invested on our end and (b) integrate their own research results, KNIME has been designed to allow for simple integration of additional types of data, algorithms, and data handling routines.

It has taken time to convince the community (and the actual users working on real world problems!) that KNIME is not yet another one of those cool but not really useful open source projects that will die away when the responsible Ph.D. student graduates. But in the past years, KNIME has increasingly gained traction in both:

[3] http://www.knime.org.

the academic community and real-life applications. It has been exciting to see how research in my own group started benefiting from a common underlying framework after a few years, but it has been a real thrill to see how this also scales to industrial users and the academic community recently. As so often, the value of the whole is much greater than the sum of the individual contributions.

4 Sparking Ideas

As outlined above, triggering new insights can still be considered the holy grail of data mining. Now, however, we understand much better what this really means. Already Wilhelm Busch knew:

> Stets findet Überraschung statt. Da, wo man's nicht erwartet hat.

Translated roughly as: surprise happens where you least expect it. However, many of the existing data mining methods and especially the tools deploying them to normal users do not really support this quest. Often, what is called "data fusion" essentially results in a very early requirement for information (source) selection and the use of automated mechanisms for feature selection which aim to minimize a predefined metric which stears that process. However, in reality, we very often do not know which data source or which metric are most appropriate to guide us toward finding the unexpected. Prediction or some other measure of accuracy on a subsample of the available data cannot be the guiding force here.

Being able to push the information source and feature selection process as deeply into the analysis process as possible will be a requirement to actually support the user in making new discoveries. Ultimately, one should be able to keep all of the available data sources involved, not only the ones that one believes from the start could have some value. Truly novel discoveries will come from discovering connections between previously unconnected domains.

Our recently concluded European Project under the acronym BISON [7] launched a first attempt to address this issue. During project definition, we stumbled across Arthur Kostler's seminal work on creativity[4] where he defined the term *bisociation* as a description of a discovery that crosses domains. This term emphasizes the difference to a typical association which is defined within a domain (or at least a well-defined union of domains). The term *Bisociative Knowlegde Discovery* therefore nicely illustrates what we are after, contrasting this against methods for well-defined pattern discovery. The project did not find the ultimate solution for this daunting task, of course, but a number of very promising directions for future research, most notably in the discovery of bridging concepts of various types, were initiated. Lots of work still needs to be done before we can actually give

[4] Arthur Koestler: *The Act of Creation*, 1964.

users a system that does support the discovery of bisociations, but we took a big step into that direction.

In my opinion, true discoveries will arise when we combine sophisticated exploration with complex, heterogeneous data sources and a multitude of data mining algorithms. The resulting systems will support the discovery of new insights across domains—and require serious efforts on joint research among the various disciplines of data mining in conjunction with researchers in the applied domains. After more than 20 years, data mining has just started to be exciting all over again!

5 Conclusions

My discussions above are naturally highly biased by the challenges we face in our daily work with researchers in Chemical Biology at Konstanz University and by our interactions with biotechs and pharma companies. Research in the other areas of data mining, on methods and theories, is, of course, still of high importance and will continue to impact research in the other areas. However, I believe that in terms of grand challenges, the development of a complex data mining system that enables a true, domain bridging discovery has the potential of a huge impact on how scientific research will be done in the future.

References

1. Michael R. Berthold, Jay Diamond, in *Boosting the Performance of RBF Networks with Dynamic Decay Adjustment*. Advances in Neural Information Processing Systems, vol 7 (MIT, Cambridge, MA, 1995), pp. 521–528
2. Rosaria Silipo, Michael R. Berthold, Input features' impact on fuzzy decision processes. IEEE Trans. Syst. Man Cybern. B **30**(6), 821–834 (2000)
3. Christian Borgelt, Michael R. Berthold, in *Mining Molecular Fragments: Finding Relevant Substructures of Molecules*. Proceedings of the IEEE International Conference on Data Mining ICDM (IEEE, 2002), pp. 51–58
4. Michael R. Berthold, in *Data Analysis in the Life Sciences: Sparking Ideas*. Knowledge Discovery in Databases, Machine Learning: PKDD/ECML 2005, Lecture Notes in AI, no. 3720 (Springer, 2005), p. 1
5. Michael R. Berthold (ed.), *Bisociative Knowledge Discovery*. Lecture Notes in Computer Science, Springer, in press
6. N. Cebron, M.R. Berthold, Active learning for object classification. Data Min. Knowl. Discov. **18**(2), 283–299 (2009)
7. Michael R. Berthold, Fabian Dill, Tobias Kötter, Kilian Thiel, in *Supporting Creativity: Towards Associative Discovery of New Insights*. Proceedings of PAKDD 2008, LNCS 5012 (Springer, 2008), pp. 14–25
8. Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, Bernd Wiswedel, in *KNIME: The Konstanz Information Miner*. Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization (Springer, 2007), pp. 319–326