On the search for and appreciation of unexpected results in data mining research (or: Science - we might be doing it wrong)

Albrecht Zimmermann

KU Leuven, Belgium

Abstract. An integral part of scientific research is the constant search for new results related to existing hypothesis, to either bolster their claims, or falsify them to advance the field. Using the example of four studies that did just that in data mining research, I will argue that the data mining community is neither interested in such studies, nor appreciates their unexpected results. Since it is my opinion that this is an attitude that holds the field back, I propose a change to the conference format that can be expected to motivate researchers to undertake more such studies, and give them higher visibility.

1 Introduction

The majority of our progress in understanding the physical world in the last 150 years, and the technological advances arising from it, can be traced back to the diligent application of the scientific method. Fields that have rejected or misapplied the scientific method, on the other hand, can be seen to stagnate or even regress. An important part of the scientific method is the repeated attempt to falsify hypotheses, i.e. to generate unexpected results, as these results lead to further progress.

As I will argue, data mining (and machine learning) research currently misapplies or even ignores the scientific method to a certain degree. Specifically, I will show that few attempts are made to systematically generate additional results related to existing work, and therefore to try and generate unexpected results. Studies that *do* generate such results are often marginalized, and their results ignored.

This article itself is admittedly not the result of a careful meta-analysis but based on personal experiences and impressions, i.e. anecdotes. I will buttress my claims by referring to more objectively accessible measures of the correctness of my claims. Specifically, I will discuss four empirical studies that challenge claims in the data mining literature and show that they were arguably not appreciated upon submission, have been ignored by the community to varying degrees, and have had their lessons ignored. Due to this subjective, and, given my research

area and experiences, arguably also somewhat myopic content, I forwent the use of the typically used communal "we" in favor of the less general "I".[1]

2 The scientific method (in the empirical sciences)

In its most simplified manner, the scientific method in the empirical sciences can be summarized in the following way:

1. The researcher has a hypothesis about the world.
2. She/he uses this hypothesis to generate a prediction.
3. He/she performs an experiment testing the prediction:
   (a) The prediction is confirmed: this is considered evidence for the hypothesis to be true, strengthening it.
   (b) The prediction is rejected: this is evidence for the hypothesis to be false, which means it has to be rejected and/or modified.

The reader will notice that this scheme implies that no matter the amount of evidence in its favor, a single (valid) counterexample is enough to falsify and reject a hypothesis – scientific hypotheses can only ever be provisionally true. The true value of negative, or in the parlance of this workshop, unexpected results lies in the rejection of hypotheses, and the resulting need for modification. A hypothesis that has generated a large amount of confirmed predictions, on the other hand, gives anyone employing it high confidence that it is, in fact, true. Unfortunately, a hypothesis that has remained unfalsified for a period of time, even if not evidence in its favor has been collected, can be mistaken for a high-confidence hypothesis as well.

As I have stated above, this is a very simplified summarization of the scientific method. For one thing, empirical experiments often do not give clear-cut *binary* results. Instead, confidence intervals are employed, likelihoods calculated, and statistical tests used to assess significance. Such assessments are more robust if experiments are *independent* from each other, ideally not only independent in time and used data but also performed by *different researchers*. Researchers, being human beings, can have biases or make mistakes and a hypothesis is the more reliable, the more different researchers have confirmed its predictions. Finally, *prior plausibility* can inform the interpretation of experimental results: a result that is barely significant and derives from a hypothesis with low prior plausibility is more likely to be accidental than a similar result stemming from a high plausibility hypothesis.

These aspects imply something that is technically not part of the scientific method: that experiments should be repeated, ideally often, ideally using equivalent but different settings, ideally by different researchers or research groups. It is this aspect of scientific work that is missing in current data mining research as I will argue in this article.

[1] The alternative option of referring to myself in the third person as "the author" felt too pompous.

3 The applicability of the scientific method to computer science

Even though computer science has "science" as part of its name, the scientific method as practiced in the empirical sciences is not simply applied as is. The main reason for this is that computer science is an *applied* science: instead of identifying new facets of how the physical world works, it uses such knowledge to build what in essence are tools for helping humans sense, process, or manipulate the world. A particular microchip is as much a tool as is a complete computer architecture, a theoretical data mining algorithm, or its optimized implementation.[2] As a side-effect of this, there is essentially only a single hypothesis in much of computer science: "it helps solve the problem". If empirical results show that the method does not help solve the problem, the hypothesis, and with it this particular design, is rejected and/or modified. But if a method solves a problem at all, no matter how inefficiently or ineffectively, a design is not rejected but added to the store of knowledge computer science has built. The question then becomes one of usefulness, i.e. how quickly the method finds solutions, and of what quality they are, and of how well assumptions about the data that motivated algorithmic design are borne out by reality. After all, the "no free lunch theorem" reminds us that there is no single method that can be expected to outperform all others over the range of all possible problems.

This is therefore where the scientific methods should find itself applied in data mining: It can be assumed that a researcher has already performed the rejection test for a new method he or she intends to propose and that methods that do not solve the problem at all will therefore not be submitted for publication.[3] Therefore, empirical evaluations should help with establishing

1. how a new method compares with the state of the art, and with similar methods.
2. what the effects of different parameter settings are on the performance of the method.[4]
3. how the data can be characterized that the method performs "well" on and how the data can be characterized on which it does not (with thanks to Eyke Hüllermeier for pointing out this blind spot of mine).

A proper data mining paper introducing a new method (or an improvement to an existing one) would therefore lay out the reasoning behind the development

[2] Since my argument here focuses on data mining, I will stop writing about tools now and in the rest of this article refer to "methods", encompassing data mining algorithms, feature selection approaches, distance measures etc.

[3] And it can be argued that this is problematic on its own: if a method with high plausibility fails to solve the problem, this is important knowledge. This argument would far exceed the scope of the article, though.

[4] This is particularly important given the "standard settings" used in toolkits such as WEKA [1] that are the main resources of many researchers that aim to employ existing methods

of the method, establishing plausibility, describe the method itself and then perform an extensive empirical evaluation. As part of this evaluation, the researcher would

- select (or generate) data covering a wide range of different data characteristics,
- identify methods most closely related (which should be easy given the plausibility analysis) as well as a number of state-of-the-art techniques that have been shown to perform well on this problem in past work,
- perform the experiments exploring a wide range of parameter settings while making sure to choose well-performing settings for the comparison techniques,
- and evaluate the results using statistical techniques to establish significance, breaking the data up into subcategories on which the method's behavior is different, i.e. where unexpected results occur.

Most of the papers I have been assigned as a reviewer fail at covering some or all these aspects. It is of course possible that those works are the unfortunate exception but if some papers that have been published in conference proceedings in past years are an indication, they are not.

Personal anecdote 1 *This has in fact become my primary rejection criterion: whether the work presents an adequate empirical evaluation of its method (where appropriate), and so far I have identified three main violations:*

1. *flaws in evaluation design: for instance by claiming to compare against the start-of-the-art but ignoring the work of the last several years, designing a worst-case strawman algorithm as the only comparison technique, or limiting the evaluation to a single data set, maybe one that is well-known to be abnormal to boot.*
2. *flaws in evaluation reporting: for instance by showing averaged accuracies without including standard deviations and/or discussing statistical significance.*
3. *overselling: for instance showing results that place the proposed method roughly equal to comparison techniques on half of the data and non-significantly better on the other half and claiming that it "significantly outperforms" the comparison techniques.*

The distressing part for me is that sometimes the argument for plausibility has been convincing enough that I would be willing to accept a paper with such a flawed evaluation if *I trusted the rest of the community to perform additional evaluations and paint a fuller picture.*

Even if a new method (or an improvement of an existing one) has been evaluated properly in the work that proposed it, it will still be desirable and necessary that it were reevaluated by other researchers, using newly surfaced data, but also using data on which it has been evaluated before, using different implementations, different experimental setups, e.g. a different number of folds, such as has

been done in the Frequent Itemset Mining Implementation competitions [2, 3]. For such studies and the possibly unexpected results they generate to have a positive impact on the community, it is necessary to give them a central spot in the literature. As I'll argue in the next section, it is in this that data mining research fails worst.

4 A number of studies showing unexpected results and their reception

In the following, I will present and briefly discuss several papers that produced unexpected results, showing established provisional truths in data mining research to be false. I will follow this up by discussing how they have been received by the community and what their impact has been as can be inferred from the literature. I want to reiterate that given the breadth of the field this is necessarily a subjective collection and subjective interpretation. I will attempt to support my claims with more objective measures, however.

Real world performance of association rule algorithms.[4] The arguably seminal itemset mining paper [5] also introduced a data generator for evaluating the running times of itemset mining algorithms. Zheng *et al.* in 2001 showed that the data generated in this manner showed different characteristics from real-life data and that the run time behavior of several algorithms [5–9] differed between the artificial and real-life data.

Implied, even though not contained, in that paper are two additional observations: First, the authors of [5] had varied the parameters of their generator, albeit in a restricted manner, to evaluate their approach on more than twenty data sets. Subsequent work in the field, however, used fewer and fewer data sets. In fact, looking at the itemset mining literature, I find an inverse correlation between the age of the paper and the number of data sets used. Second, Zheng *et al.* showed that CHARM significantly outperformed CLOSET on the real-life data, a finding that contradicted the results reported in [8], in which different data had been used. As Zaki then showed in [6], CHARM also outperformed CLOSET on the data used [8] if one lowered minimum support further than had been done in that work. The differences in performance can therefore be tied both to data and parameter settings.

Using Classification to Evaluate the Output of Confidence-Based Association Rule Mining.[10] Association-based classification had first been proposed in [11]. Mutter *et al.* showed in their work that CBA, the algorithm proposed in [11], did not perform better than existing rule-based classifiers, putting the results of that work into perspective and partially contradicting them. Their work furthermore showed that replacing APRIORI by PREDICTIVE APRIORI proposed by Scheffer [12] lead to better classifiers.

As a side note, the authors write in their paper that they were not able to reproduce all the results from [11] with a reimplementation of their own, and

report in the thesis that the paper was based on that they also could not achieve this with an implementation provided by the authors of [11].

Personal anecdote 2 *While working on a past paper, we planned to compare to a technique proposed by another data mining researcher. We obtained both the original data and the implementation from the author but did not manage to reproduce some results. After mailing him about the issue, we received a reply along the lines that he did not understand our problem since he had been able to reproduce the results. The result files were attached. The solution to the problem was that while the results from the paper could be obtained, they could not be obtained using the parameter settings from which the paper claimed they originated – apparently the parameter entries had been switched around on writing the paper.*

Obtaining Best Parameter Values for Accurate Classification.[13] Association based classification usually works with standard values with 1% for minimum support and 50% for minimum confidence. As Coenen *et al.* showed experimentally, these values are not only not the best values for achieving high accuracy but especially CBA often tended to perform worst for these settings.

Frequent Subgraph Miners: Runtimes don't Say Everything.[14] Improving running times is a major goal in the development of new pattern mining algorithms. As Nijssen *et al.* showed, many of the claimed underlying reasons for improved run times did not hold up under scrutiny. Additionally, they found that the interplay between cache size of the processor used and the size of the data set, i.e. aspects largely outside of researchers' control, had a strong effect.

4.1 Reception in the community

Given the works I have just listed, one could be tempted to assume that all is well in data mining research since such works are being undertaken and papers based on them published. The reader could also be forgiven for pointing towards the large scale evaluations the group of Johannes Fürnkranz has undertaken over the years, for instance, on the effects of coverage and consistency in rule-learning heuristic [15], *beam sizes* [16], probability estimation techniques [17] etc and absolving machine learning research.

The problem, however, is not so much with whether these works are done in the first place – even though a cursory glance through a year's data mining conferences will show far more papers proposing new methods than ones further evaluating existing ones. In my opinion, the problem lies instead with the effect (or lack thereof) such studies have on the field. It is difficult to evaluate this effect directly which is why I will use three proxies instead:

1. the venue they were published in, and whether they were full papers.
2. the number of citations (related to the number of citations the papers proposing the evaluated methods received).
3. the effect as can be inferred from the literature.

Venues and paper type Zheng *et al.* [4]: short paper at KDD 2001. Mutter *et al.* [10]: rejected at ECML/PKDD, regular paper at the Australian Joint Conference on Artificial Intelligence 2004, which is not highly ranked according to different rankings. Coenen *et al.* [13]: short paper at ICDM 2005, which did not have short paper submissions. Nijssen *et al.* [14]: workshop paper, although Siegfried Nijssen claims that he simply never followed up on this work.[5]

The situation for the Fürnkranz publications is similar: [18] short paper ICDM 2007, [15] Discovery Science 2008, not a highly ranked conference, [16] poster SDM 2009, [17] Discovery Science.

These studies should have been front-page news, for challenging provisional truths and showing that accepted default values and default experimental setups were faulty. Instead there were shunted aside in favor of yet another weakly understood new method. To be clear: there is nothing wrong with publishing at conferences that are not highly ranked. It just indicates that the community is not interested enough in these works to see them published in full at the conferences it values highly.[6] This makes it less likely that these works are noticed by the wider community and the lessons learned incorporated. This has not only to do with those works' appearance in the conferences' proceedings but at least as much with having the opportunity of giving a talk to a large and interested audience and impressing the importance of the issue on them. KDD 2011 had more than 1000 attendants. ICDM 2011, while still being a high-ranked conference, had in the range of 400-600. Conferences like Discovery Science are more likely to attract in the vicinity of 200-300.

Even if such a study *is* included in a high-ranking conference, the nature of such evaluations is that quite a few different experimental settings are used and lots of numbers produced. If a write-up is then forced to do with a reduced page limit, it will be difficult to present the studies and their results comprehensively.

Citation count I used Google Scholar on August 5th, 2012, with all that this entails, and compare each paper's citation count with that of the methods evaluated. If the insights derived from these studies found application in data mining research and default values or experimental setups were adjusted accordingly, I would expect a citation count for these studies that at least comes close to those of the methods they correct.

Zheng *et al.*: 352 citations, compared to 13020, 990, 4178, 779, 149 for the evaluated methods (in order of citation above).

Mutter *et al.*: 17 citations, 3 by papers of which I am a co-author, 2 by papers of which Johannes Fürnkranz is a co-author, Coenen *et al.*: 26 citations, 7 self, 2 by paper of which I am a co-author, compared to 1582 (CBA), and 920 (CMAR) citations.

Nijssen *et al.*: 16 citations, compared to 1035 (gSpan), 346 (FFSM), 67 (AcGM), 779 (FSG).

[5] personal communication

[6] Why these conferences are highly ranked is another question to do with publishing arcana.

The Fürnkranz papers: [18] 14 citations, 7 self, [15] 9 citations, 6 self, [16] 7 citations, 5 self, 2 by papers of which I was co-author, [17] 3 citations, 2 self.

Impact on the field Zheng *et al.* certainly had an impact since the data sets they introduced have become standard benchmark sets and the data generator proposed in [5] has fallen into disregard. However, the deeper lesson, that focusing on a small number of data sets in proposing improvements can lead to overfitting effects, has clearly been lost on the community given the small number of data sets used to evaluate itemset mining approaches, the lack of a widely-used replacement generator, and the few large-scale comparisons [2, 3].

Given the results of Mutter *et al.* and Coenen *et al.*, the experimental evaluations of associative classification research since 2005 should have featured the use of PREDICTIVE APRIORI and at least an exploration of a range of support-confidence combinations. A survey of the relevant literature since then will disabuse the reader of this notion.

The insights derived in Nijssen *et al.*, while formulated in the context of graph mining approaches, should be heeded in all pattern mining research but claims of speed-ups having to do with canonical forms still abound.

4.2 Is this really a problem?

The reader can now (and maybe already has) argue that this is not a real problem. After all, three of the studies I discussed are concerned with itemset mining, a field that has been researched for a while and in which no groundbreaking discoveries can be expected anymore. I would have to disagree since there is no reason to assume that the scientific method is applied more diligently in other fields as the example of Nijssens *et al.* shows. Even if this were the case, data mining research, as I have argued in the beginning, is essentially concerned with building tools, such as itemset mining techniques. If these tools are never used outside of academia because users do not know under which conditions they might be useful, data mining research misses its purpose.

The reader might claim that data mining (or the subfields of data mining I am most familiar with) is an aberration and the scientific method is employed much more stringently in other areas. However, the examples of the works from the group of Johannes Fürnkranz lead me to believe that those problems exist in the machine learning community as well. But even if this is not the case, it seems even more urgent to change the practice in data mining research lest the area stagnate.

Finally, the reader could argue that the methods that truly make an impact will be evaluated, and their working parameters established, in successive studies as they are used more and more often. This has, for instance, happened to decision tree or support vector machine classifiers. Non-performing methods get drowned by the tide of papers published every year and so no further evaluation is necessary.

There are at least three problems with this idea: the first is that in the absence of principled studies we do not know that the methods that rise to the

top are truly the best (or even reasonably good) ones. They might instead be the early ones that did not have to compete with many other papers yet, they might be the simplest ones that everyone understands, or they might originate from a large research group whose cross-citations help them gain critical mass. The second problem is, as can be seen in itemset mining, that those principled evaluations might simply never happen, even for well-known techniques. Third, the proliferation of methods is intrinsic to the issue I am discussing here: the relative disregard shown to works searching for unexpected results makes it appear more attractive to propose new methods (with "good" results) instead, and the lack of scientific evaluations translates into a lack of guidance regarding the methods that should be improved. This also means that lots of time, money, and energy are wasted on dead end research that does not improve the field.

Furthermore, the issues I have described so far are connected to relatively clear-cut evaluation criteria: running times and classification accuracy. In descriptive data mining, which includes most of pattern mining, such clear-cut evaluation criteria are not available. Instead, the problem to be solved there consists of extracting underlying patterns of correlations in the data.

Remarkably enough, in at least two subfields, itemset mining and frequent episode mining, the literature so far does not include any evaluations on whether extracted patterns correspond to known phenomena in the data. Instead, attempts have been made to evaluate the quality of found patterns by presenting them to domain experts who were supposed to perform this evaluation [19]. An implicit assumption in these kinds of evaluations is that the experts will be able to properly identify interesting patterns as interesting and uninteresting ones as uninteresting. The plausibility of this assumption is however unknown and if psychological research on humans' tendency to see patterns is any guide, it might be much lower than the authors of such studies assume.

An experiment that I would like to see would consist of calibrating the domain experts first: instead of having them evaluate significant patterns, they would be given a mix of significant, non-significant, and random patterns and their "accuracy" evaluated.

Personal anecdote 3 *Last year, I had what I considered to be a "small" idea for a conference paper. It had to do with feature generation, had good plausibility, discussions with colleagues revealed no obvious flaws, and the literature study showed that it had not been tried before. So I went for it - and the results were atrocious. Assuming a mistake on my side, I checked everything several times but could not find one. Finally, I gave up and tried to get this negative result published, as a warning to others. The submission was rejected with one reviewer writing that the results did not surprise him (and the two others that they did not see the use of publishing a negative result until everything had been tried to make the approach work). The reviewer admitted that she/he could not provide a reference but provided a convincing rationalization.*

Long story short, when Joaquin Vanschoren asked me to perform some further analysis on my results for the workshop, I revisited the old scripts - and found the bug that had escaped me. With the bug corrected, the approach works as I

expected. Unless the reviewer knows more about my code-writing prowess than I do (and just wanted to spare my feelings), his/her lack of surprise is surprising.

If I had managed to get these false results published, I would not have revisited them, and since the results were negative, in all likelihood no one else in the community would have.

5 What we can do to get back on track

In my opinion, there are at least two causes that can be identified underlying the attitude of the data mining and machine learning community towards generating, reporting and using unexpected results, one that is for now outside of our power to effectively address, but also one that could be addressed by small changes to the current conference format.

First, it is my impression that the scientific method, and specifically the need for constant retesting, and the importance and usefulness of unexpected results, is not being impressed on young researchers. This has ripple effects, leading to the aforementioned badly designed (and reported) experimental results, an unwillingness to report unexpected results, and a tendency to negatively review papers that do report new results of existing techniques or unexpected (negative) results. There is not much we, as individuals, can do to change this: we can try and influence students and colleagues, we can criticize weak experimental evaluations in reviews and point out avenues for improvement, we can support studies searching for unexpected results. But all of these can only be expected to be drops in the ocean given the amount of new researchers and new publications each year.

Second, purely experimental studies that do not propose a new method do not have a "home" in the community's conferences. The typical conference format in 2012 was to have a "research" track and either no (SDM, CIKM, ICML, ECAI, DS), or only one (KDD, ICDM, CIKM, ICDE) additional track (e.g. industry and government) for submission (ECML/PKDD being the exception with two extra tracks). The calls for papers make it explicit that this is intended in listing the kind of papers that are solicited and papers are supposed to be subdivided by using keywords during submission, often ones focused on problem areas. This means that at least four different types of papers, all of which are subject to the same page restrictions, compete for acceptance:

1. Papers proposing new methods: these works have to establish plausibility, outline the method, and ideally show an extensive experimental evaluation. "Good" results on at least some data will be helpful in establishing the usefulness of the new method.
2. Papers proposing improvements to existing methods: since the main method has already been introduced, these works only have to establish plausibility of (and describe) the proposed improvement, and the experimental evaluation can focus on showing the effects of the proposal (but should still do so thoroughly). These effects arguably need to be positive for it to be an improvement.

3. Experimental evaluations of existing methods: these works need to establish the appropriateness and/or difference of their data and experimental settings for producing new results, and need to present and analyze their experimental results in detail.
4. Theoretical works: are something rather different.

The "success" criteria are thus different for all four types, with some easier to evaluate than others, yet there is typically a one-size-fits-none reviewing framework within which they are supposed to be evaluated.

The reader might think that journals, with their larger page count per paper and specifically selected reviewers, can be an alternative to conferences but those also have an attention problem: an extended version of a paper introducing a new method will necessarily appear later than its conference version so that a researcher who has read the conference version might not go to the effort of reading the journal version as well. Instead of a program that every conference participant is handed, the table of contents of a journal must be actively sought out by researchers, and a journal paper is not accompanied by a presentation. Finally, that a journal paper is longer can paradoxically work to its disadvantage since a researcher seeking a quick understanding of the method would likely prefer the shorter conference version.[7]

On the other hand, given the current use of keywords to subdivide submissions by areas but also to a certain degree by content, e.g. foundations of data mining, it should easily be possible to invite submissions to several different tracks. This can be expected to motivate researchers on the fence about attempting (and attempting to publish) works generating additional results for existing methods, and it should allow better-targeted instructions to the reviewers.

Outside of the publication context, there already exist projects that collect experimental results for different data sets, algorithms, and parameter settings, such as ExpDB[8] and MLComp[9] to enable comparison and the selection of appropriate methods. These projects cannot be an alternative to a rethinking of the scientific process in data mining research and supply a clean conscience, however. Instead, they offer support for comparison and their success *depends* on a change of mind among data mining researchers.

6 Conclusion

In this article, I have discussed the application of the scientific method in data mining research, specifically the search for and appreciation of unexpected results, and found the field lacking in this regard. A study of a typical year's literature reveals a plethora of new and improved methods and few systematic evaluations of these methods. If such evaluations are done at all, they often show

[7] Additionally, journals also have the competition issue, albeit to a lesser degree.
[8] http://expdb.cs.kuleuven.be/expdb/
[9] http://mlcomp.org/

well-established provisional truths to be wrong, as I have argued using the example of four empirical studies. However, since they tend to get marginalized during publication, the derived insights do not find their expression in future research, and those provisional truths (while shown to be false) continue to be accepted.

In my opinion, this self-perpetuating cycle can be attacked by acknowledging that there is more to data mining research than just the proposal of yet another method, and by changing the conference format to motivate researchers to undertake the work necessary for understanding the strengths and limitations of existing data mining methods. The goal of data mining research lies in providing tools to users. This does not mean, however, that there is any merit in cranking out ever more methods without learning under what conditions those methods work well. The current pool of often-used methods is limited to the few that are, if not well-understood, at least well-known and unless we change the way data mining research is conducted, it will stay like this.

Acknowledgements

The motivation for writing this article and quite a few of the arguments laid out stem mainly from discussions with numerous participants of the FLF workshops, too many to mention them all here. I would like to single out Matthijs van Leeuwen for giving important feedback on a first draft on short notice, and Tias Guns who, even though he did not any have time for this, offered up several of the counter arguments.

References

1. Frank, E., Witten, I.H.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999)
2. Goethals, B., Zaki, M.J., eds.: FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA. In Goethals, B., Zaki, M.J., eds.: FIMI. Volume 90 of CEUR Workshop Proceedings., CEUR-WS.org (2003)
3. Bayardo Jr., R.J., Goethals, B., Zaki, M.J., eds.: FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1, 2004. In Bayardo Jr., R.J., Goethals, B., Zaki, M.J., eds.: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations. (2004)
4. Zheng, Z., Kohavi, R., Mason, L.: Real world performance of association rule algorithms. In: KDD. (2001) 401–406
5. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Databases, Santiago de Chile, Chile, Morgan Kaufmann (September 1994) 487–499
6. Zaki, M.J., Hsiao, C.J.: Charm: An efficient algorithm for closed itemset mining. In Grossman, R.L., Han, J., Kumar, V., Mannila, H., Motwani, R., eds.: SDM, SIAM (2002)

7. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas,Texas,USA, ACM (May 2000) 1–12
8. Pei, J., Han, J., Mao, R.: Closet: An efficient algorithm for mining frequent closed itemsets. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. (2000) 21–30
9. Webb, G.I.: Efficient search for association rules. In: KDD. (2000) 99–107
10. Mutter, S., Hall, M., Frank, E.: Using classification to evaluate the output of confidence-based association rule mining. In Webb, G.I., Yu, X., eds.: Proceedings of the 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, Springer (December 2004) 538–549
11. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In Agrawal, R., Stolorz, P.E., Piatetsky-Shapiro, G., eds.: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York City, New York, USA, AAAI Press (August 1998) 80–86
12. Scheffer, T.: Finding association rules that trade support optimally against confidence. In De Raedt, L., Siebes, A., eds.: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, Freiburg,Germany, Springer (September 2001) 424–435
13. Coenen, F., Leng, P.: Obtaining best parameter values for accurate classification. In Han, J., Wah, B.W., Raghavan, V., Wu, X., Rastogi, R., eds.: Proceedings of the Fifth IEEE International Conference on Data Mining, Houston, Texas, USA, IEEE (November 2005) 597–600
14. Nijssen, S., Kok, J.: Frequent subgraph miners: runtimes don't say everything. In Gärtner, T., Garriga, G., Meinl, T., eds.: Proceedings of the Workshop on Mining and Learning with Graphs,. (2006) 173–180
15. Janssen, F., Fürnkranz, J.: An empirical investigation of the trade-off between consistency and coverage in rule learning heuristics. In Boulicaut, J.F., Berthold, M.R., Horváth, T., eds.: Discovery Science. Volume 5255 of Lecture Notes in Computer Science., Springer (2008) 40–51
16. Janssen, F., Fürnkranz, J.: A re-evaluation of the over-searching phenomenon in inductive rule learning. In: SDM, SIAM (2009) 329–340
17. Sulzmann, J.N., Fürnkranz, J.: An empirical comparison of probability estimation techniques for probabilistic rules. In Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P., eds.: Discovery Science. Volume 5808 of Lecture Notes in Computer Science., Springer (2009) 317–331
18. Janssen, F., Fürnkranz, J.: On meta-learning rule learning heuristics. In Ramakrishnan, N., Zaiane, O., eds.: ICDM, IEEE Computer Society (2007) 529–534
19. Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In Boulicaut, J.F., Esposito, F., Giannotti, F., Pedreschi, D., eds.: PKDD. Volume 3202 of Lecture Notes in Computer Science., Springer (2004) 362–373