On the Importance of Sharing Negative Results

Christophe Giraud-Carrier
Department of Computer Science
Brigham Young University
Provo, UT 84602, USA
cgc@cs.byu.edu

Margaret H. Dunham
Computer Science and Engineering Department
Southern Methodist University
Dallas, TX 75275, USA
mhd@lyle.smu.edu

1. INTRODUCTION

The empirical study of machine learning and data mining methods often falls prey to the effects of publication bias that favors positive results over negative ones. Most, if not all, articles in conferences and journals report only positive results. This does not reflect the practice of a field where failures happen regularly. As in real life, we often learn more from negative results than we do from positive ones. It is time that we, as a community, start to regard failures as being as informative as successes. After all, we do know the difficulty of learning from positive only experiences; so how can we expect to learn about our field if all we ever see are successes?

This special issue provides a forum for papers that describe clear, and somehow surprising, failures that stand in need of an explanation. We define as clear, or interesting, failures that happen in situations where the learning or mining method is not only sub-optimal, but performs far worse than expected. To make the special issue of value to the largest possible audience, we sought papers that report failures of learning and mining strategies that are already popular and well-known in the community, or of novel ideas that are easy to comprehend and do not require extensive prior knowledge in a special niche area of machine learning or data mining. The main purpose of this special issue therefore is to bring together a sample of exemplary failures, with the goals of:

1. making these experiences accessible to fellow researchers who may otherwise waste their time on the same or similar idea, and

2. documenting the first few negative data points necessary to gain additional insights into our methods (e.g., what method is applicable where).

We are aware of only one other previous collection of "unexpected results" articles devoted to data mining. A special issue of *Machine Learning* (Vol. 57, 2004) on Data Mining Lessons Learned —initiated following an ICML-2002 workshop of the same name— started with a premise similar to ours and included 6 contributed papers. These were complete papers whose focus was generally: "here is the challenge we faced and here is how we overcame it." In this special issue, we wish to consider more of the situations of: "here is what we thought would happen and here is what actually happened." The contributed papers are intentionally short, speaking directly to the negative and/or unexpected nature of the reported results.

2. THE NEED FOR SHARING NEGATIVE RESULTS

The importance of negative results has been recognized in several scientific disciplines, as evidenced by well-exploited peer-reviewed publications dedicated to such results. We list a few here as illustration.

- The *Journal of Negative Results in Biomedicine* is an online journal dedicated to the "discussion of unexpected, controversial, provocative and/or or negative results in the context of current tenets" in biomedicine.[1]

- The *All Results Journals* "focuses on recovering and publishing negative results, valuable pieces of information... considered a vital key for the development of science and the catalyst for a real science-based empirical knowledge" in nanotechnology, chemistry, biology, and physics.[2]

- The *Journal of Negative Results* offers a forum for "the publication of... sound scientific work in ecology and evolutionary biology that may otherwise remain unknown," with a special focus on "studies that seem uneventful," to counteract what may otherwise "lead to a biased, perhaps untrue, representation of what exists in nature."[3]

Interestingly, machine learning and data mining too have a journal. This may come as a surprise to some readers. It is called the *Journal of Interesting Negative Results*, and its explicit aim is to be "a resource that gives a voice to negative results which stem from intuitive and justifiable ideas, proven wrong through thorough and well-conducted experiments...[as well as] short papers/communications presenting counter-examples to usually accepted conjectures or to published papers."[4] The journal started in April 2008, but since that time has published only one article!

If other disciplines recognize the importance of negative results to advancing knowledge about their disciplines why don't we (Computer Scientists)? While we can't answer this conclusively, we think some of it may have to do with

[1] http://www.jnrbm.com/
[2] http://www.arjournals.com/ojs/
[3] http://www.jnr-eeb.org/
[4] http://jinr.site.uottawa.ca/

the fact that Computer Science grew out of mathematics as opposed to a more clinical or laboratory based science. However, over the past twenty years Computer Science has moved into a very much more applied discipline and perhaps the Data Mining sub-discipline is leading this evolution. In fact, since most of the problems we deal with are interdisciplinary, and require the design and evaluation of experiments perhaps we, as data miners, have much more to learn from negative results. After all, we are not mathematicians any more!

The scientific method itself allows for failure —we need to accept that negative results are part of our everyday professional life. The metalearning that occurs through evaluation and reflection should be a valuable and important part of our work.

3. SUMMARY OF THIS ISSUE

The papers published in this special issue highlight unexpected results found in data mining experiments. These results can be summarized into lessons learned about the data mining technique, handling of data, and the importance of careful design of experiments.

3.1 Not Every Data Mining Technique Is Applicable in All Situations

Atreya and Elkan report that Latent Semantic Indexing (LSI), a popular method for text analysis, performs very poorly on several of the benchmark TREC document collections. Despite trying several versions, no version of LSI achieves a worthwhile improvement in retrieval accuracy over BM25, the best currently known vector-based scoring method. It is hypothesized that the reason may be the large number of dimensions in the problem. However, experiments conducted do not validate this hypothesis. The authors have not yet been able to conclusively determine the reason for this poor performance of LSI.

Perlich and Świrszcz report on some surprising results when applying cross validation to derive conclusions about data especially when the (positive) signal is weak. The authors suggest that the technique generally produces an inverse signal, which in turn, yields extremely low AUC (Area Under the Curve) values. The authors show that the problem particularly affects popular ensemble methods, such as bagging, which are commonly regarded as very robust.

Shi and Yu explore the limitations of trace norm minimization, particularly as a way of replacing missing values in a matrix (e.g., as is necessary for collaborative filtering). They point out that the main assumption of this approach, namely that the original matrix is of low rank, cannot be verified in practice. In addition, it may produce multiple solutions each of the same low rank. The authors conclude that trace norm minimization thus only works under certain very constrained situations.

3.2 Beware of Choice and Use of Input Data

Fürnkranz and Sima report on unexpected results of experiments with multilabel data mining where input data is augmented with information about the hierarchical relationships among the objects. The authors show that for binary class hierarchies this does the same as the well-known Pachinko machine. However it trains many redundant and therefore useless classifiers. It also performs worse than the normal not augmented pairwise classification. The problem seems to lie in the fact that the evaluation domain does not satisfy the authors' so-called class fidelity assumption, i.e. assuming that instances are closer if their classes are closer in the label-hierarchy.

Weninger et al investigate problems associated with automatically extracting lists from the Web. Contrary to their expectations, it seems that an extremely naïve approach outperforms existing more sophisticated techniques. Using the structure of the Web page and HTML clues is not sufficient to perform this task.

3.3 We Must Be More Careful in Designing Our Experiments

Kohavi and Longbotham share several unexpected and erroneous results found from experience in performing many different online randomized experiments. They point out many issues to be careful about when conducting these types of controlled experiments. Although many of the results, such as the impact of caching and redirects, are obvious once explained, they are not always the sorts of things that would be thought of when designing the experiments.

In their reflective article, Attenberg and Provost discuss a number of challenging issues or questions that must be addressed in using active learning in practice. While active learning promises to reduce the cost of acquiring labeled data, most research in the area overlooks some important issues that make it impractical, such as, how to choose a technique, how to choose a base learner, how to deal with skewed distributions and disjunctive classes, and how to "start" the process.

4. CONCLUSION

The three major themes of the included papers: techniques, data, and experiments indeed highlight the major components of our field. It is interesting to note that one of the major issues evolving in Computer Science education is that of effective data analysis. We also find it interesting that there are very few "design and analysis of experiments" classes in undergraduate Computer Science curricula. We feel strongly that our sub-discipline needs the reflective input provided by evaluating the failures of our own work. We strongly recommend that, in addition to the more formal *Journal of Interesting Negative Results*, *SIGKDD Explorations* and/or our flagship conferences (e.g., *KDD*, *ICML*, etc.) have a regular feature on carefully documented negative experimental results.

The publication of this special issue would not have been possible without the enthusiastic responses of our invited colleagues, Charles Elkan, Jiawei Han, Ronny Kohavi, Foster Provost, and Philip Yu, as well as the voluntary submissions of others, from which two papers were selected. We are also grateful to Chris Drummond, Johannes Fürnkranz, Robbie Haertel, Michael Hahsler, Jörg-Uwe Kietz, Mallik Kotamarti, Gregory Piatetsky-Shapiro, Carlos Soares and Ricardo Vilalta, for helping us with the review of the papers.