Data Mining Disasters: a report

Mary McGlohon
Carnegie Mellon University
Machine Forgetting Department
5000 Forbes Ave.
Pittsburgh, Penn. USA
mmcgloho@cs.cmu.edu

Figure 1: ERROR::NumericOverflow. Nobody anticipated the breach of the levees.

ABSTRACT

Preventing data mining disasters is an important problem in ensuring the profitability and safety of the field of data mining. Some data mining disasters include decision tree forest fires, numerical overflow, power law failure, dangerous BLASTing, and an associated risk of voting fraud. This work surveys a number of data mining disasters and proposes several prevention techniques.

1. DATA MINING DISASTERS AND RECOMMENDATIONS

1.1 Numeric overflow

Numeric overflow is a significant problem in machine learning programming. In 2007, numeric floods caused over $600 million in property damages [1], and a loss of several thousand nerd-hours of work.[1] A lack of response fromthe Programming Emergency Management Agency (PEMA) was also often cited as an issue in such catastrophes.

When faced with a situation of numeric floods (such as that shown in Fig. 1.1), a drowning researcher's best bet is to grab hold of a floating `log` among the debris.

[1] 1 nerd-hour = 1 grad-student hour = 6 undergrad-hours = 0.5 faculty-hours

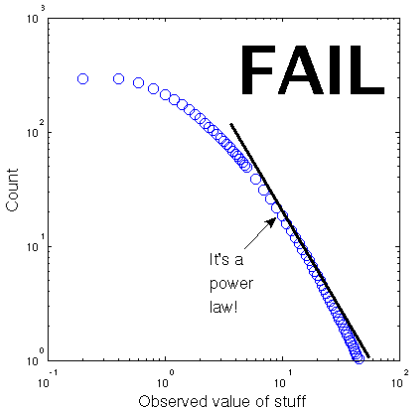

Figure 2: This is probably a log-normal distribution. This is not a power law.

1.2 Power law failures

While much natural phenomena follow long-tailed distributions, there is a tendency to believe that everything is self-similar and that all long-tailed distributions are equivalent to power-laws (see Fig. 1.2). This has become a source of debate between computer scientists, physicists, and statisticians. The last group tends to be very particular on what constitutes a "distribution". A debate may be found in [3, 9].

Techniques for avoiding this sort of power-law failure are described in detail in [4].

A possibly more dire form of power-law failure occurs when researchers spend too much time arguing whether or not some long-tailed-looking data actually comes from a power law, log-normal, or doubly-Pareto log-normal generator. Everybody knows that things get nasty when statisticians get religious about something (for instance, the turf wars between rapping statisticians Emcee M.C. and the Unbiased M.L.E [7]).

1.3 Decision tree forest fires

Occasionally researchers using pruning algorithms on their decision trees get carried away. Instead of pruning unnecessary branches in the interests of reducing overfitting. The experimenter just burns down the tree until it is a decision stump. Repeating this on every decision tree built is what is termed a *decision tree forest fire* (see Fig. 3). This is not to

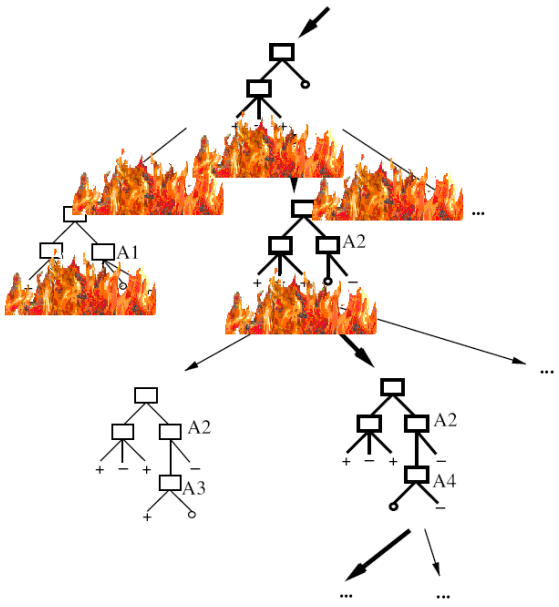

Figure 3: Remember, kids, only you can prevent decision tree forest fires.

be confused with the *Forest Fire Model*, a generative model for evolving social networks [5].

As prevention measures, researchers should obtain a burning permit before choosing to prune their decision trees with fire. Also, smoking while researching is not recommended, and anyone engaging in such behavior should ensure that their "butts are out".

1.4 BLAST accidents

Bioinformatic tool *Basic Local Alignment Search Tool (BLAST)* [2] is useful for comparing sequences of amino-acids in proteins, or of base-pairs in DNA sequences. However, if used improperly, it can be over-sensitive. This is what we term a *mining BLAST accident*.

A recommendation to avoid such disasters it for researchers to be properly trained in using BLAST, as well as alternative algorithms for subsequence matching.

1.5 Voting fraud by one-armed bandits

Data mining also may suffer cascading failures from errors made in other fields. Two important game theory and mechanism design subfields are voting mechanisms and one-armed bandit problems [10]. A fatal mistake is made when combining the two, which results in inaccurate data; thereby creating data mining disasters when data mining researchers attempt to use these data.

There are several common methods that one-armed bandits use of committing voter fraud. For instance, they may *impersonate* actual voting machines (see Fig. 4). They may also try to confuse polling officials by citing various violations of policies set by the *Americans with Disabilities Act*. They may also cram cake[6] into the voting machines[2].

[2]The cake is a lie.

Figure 4: This is what happens when you don't pay attention in your undergrad AI class.

Figure 5: Regulation safety helmets for data miners can prevent accidents.

2. OTHER PREVENTION TECHNIQUES

2.1 Cool Helmets

As a safety precaution, data miners should wear mining helmets, such as that shown in Fig. 5. And overalls, ideally. This will also serve to legitimize data mining as a real field of mining.[3] As a result, it will raise morale among researchers and prevent the often fatal results of data mining accidents.

3. CONCLUSIONS

The author hopes that this paper will raise awareness among data miners of risks involved in the field of practical prevention techniques. When faced with any sort of data mining disaster, it is generally advisable to remain calm and

[3]Talismans such as scarves, fanny packs, and pony-tails may also serve as good-luck charms in preventing data mining disasters.

blame it on one-off errors, lack of rigor in proofs of correctness, or whatever government agency is funding the project.

Acknowledgments

Some images were borrowed from various sources on the Internet and blatantly defiled with MS Paint. The original image used in Fig. 1.1 was provided by the Associated Press. The image for Fig. 3 was borrowed from Tom Mitchell's webpage for his textbook [8]. Sources for Fig. 4 include `digitalmedia.ucf.edu` and `www.thewe.cc`. In Fig. 5, Christos Faloutsos is modeling a mining helmet found at `goldenwesttravel.net`.

4. REFERENCES

[1] Made up statistics, 2008.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.

[3] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005.

[4] A. Clauset, C. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. 2007.

[5] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA, 2005. ACM Press.

[6] M. Magdon-Ismail, C. Busch, and M. Krishnamoorthy. Cake-cutting is not a piece of cake, 2002.

[7] M. McGlohon. Methods and uses of graph demoralization. In *The 6th Biarennial Workshop about Symposium on Robot Dance Party of Conference in Celebration of Harry Q. Bovik's 0x40th Birthday*, Apr. 2007.

[8] T. Mitchell. *Machine Learning*. McGraw-Hill Education (ISE Editions), October 1997.

[9] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral. Comment on barabasi, nature 435, 207 (2005). 2005.

[10] M. Wooldridge. *Introduction to MultiAgent Systems*. John Wiley & Sons, June 2002.