Situational Awareness at Internet Scale – Detection of Extremely Rare Crisis Periods

David Cieslak, dcieslak@cse.nd.edu, http://www.nd.edu/~dcieslak/, 8962
Philip Kegelmeyer, wpk@sandia.gov, csmr.ca.sandia.gov/~wpk 8962

It would valuable to have early warning of disruptions to the Internet. One aspect of Internet function is the Border Gateway Protocol (BGP), which handles the delivery of packets between various autonomous domains. In particular, autonomous routers use BGP communications for "announcements" and "withdrawals", that is, to announce the detected availability or unavailability of various Internet routes.

Certain statistics of these announcements and withdrawals might be useful in distinguishing between normal and abnormal operation of the Internet[3]. Practical use of these statistics in real-time monitoring is stymied, however, by the need for robust pattern recognition methods that can operate under conditions of extreme skew. In other words, the disruptions, though critical when they occur, nonetheless are occuring only a tiny fraction of the time: in the last decade of Internet activity, there has been roughly twenty cumulative days of "disruption", an incidence rate of 0.5%. This problem of imbalanced data afflicts many applications, and is becoming even more prevalent as data volumes grow.

"Bagged" ensembles of decision trees have been shown to be a simple, robust and accurate method for the detection of events in noisy data[1]. But even bagged trees falter in the face of such extreme skew. After all, if a disruption event shows up only 0.5% of the time, then a classifier could generate the trivial rule of classifying everything as "normal" and still be 99.5% accurate.

Thus we require both analysis methods *and* accuracy metrics more appropriate to skew data. Hellinger trees are a new decision tree analysis method that have been shown to be statistically significantly more accurate and robust than Infogain trees[2] (the best current decision tree method) when applied to skew data. Further, Hellinger trees have also been shown to be statistically significantly superior to pre-processing skew data with SMOTE, which was previously the best known corrective for skew data.

These advantages have been established, however, only in the context of the use of a *single* tree, not ensembles. Accordingly, we are investigating the application of ensembles of Hellinger trees to skew data, starting with the integration of Hellinger trees into the "AvatarTools"[4] tool set. AvatarTools is a Sandia developed and maintained suite of executables for supervised machine learning via ensembles of decision trees. For the current purposes, its differentiating capabilities are the ability to automatically choose the proper ensemble size, and the implementation of a slew of skew correction methods. The latter serve as a comparative baseline for Hellinger trees.

In experimentation with AvatarTools over a wide variety of test data, we have been able to show that ensembles of Hellinger trees are statistically significantly superior to

ensembles of Infogain trees on skew data, and also significantly statistically no worse than Infogain trees on balanced data. In sum, this means that anyone using decision trees should always use Hellinger trees, as they never hurt and often substantially help.

Accordingly, we are investigating the use of ensembles of Hellinger trees in the analysis of BGP data for the detection of disruptions in Internet activity. We have acquired real world BGP data[5] that captures activity before, during, and after a variety of known historical disruptions[1], including worms, fiber cuts, distributed denial of service attacks, and power blackouts. We have used this real data to construct and study skew scenarios as extreme as 1000:1. We have shown that bagged ensembles of Hellinger trees do outperform bagged ensembles of Infogain trees, and that the performance gap widens monotonically with increasing skew.

References

[1] Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 1 (January 2007), 173–180.

[2] Cieslak, D. A., and Chawla, N. V. Learning decision trees for unbalanced data. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (Antwerp, Belgium, September 2008).

[3] Dou, D., Li, J., Qin, H., Kim, S., and Zhong, S. Understanding and utilizing the hierarchy of abnormal BGP events. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (2007), pp. 467–472.

[4] Kegelmeyer, P., Buch, K., and Cieslak, D. Avatartools. `www.ca.sandia.gov/avatar`.

[5] RIPE NCC. RIPE routing information service raw data. `www.ripe.net`.

[1]We thank Max Planck of New Mexico Technical Institute for his invaluable work in acquiring, pre-processing, and truthing this data.