Data Quality Assurance and Performance Measurement of Data Mining for Preventive Maintenance of Power Grid

Leon Wu
Department of Computer Science
Columbia University
New York, NY 10027
leon@cs.columbia.edu

Gail Kaiser
Department of Computer Science
Columbia University
New York, NY 10027
kaiser@cs.columbia.edu

Cynthia Rudin
MIT Sloan School of Management
MIT
Cambridge, MA 02139
rudin@mit.edu

Roger Anderson
Center for Computational Learning Systems
Columbia University
New York, NY 10115
anderson@ccls.columbia.edu

ABSTRACT

Ensuring reliability as the electrical grid morphs into the "smart grid" will require innovations in how we assess the state of the grid, for the purpose of proactive maintenance, rather than reactive maintenance; in the future, we will not only react to failures, but also try to anticipate and avoid them using predictive modeling (machine learning and data mining) techniques. To help in meeting this challenge, we present the *Neutral Online Visualization-aided Autonomic evaluation framework* (NOVA) for evaluating machine learning and data mining algorithms for preventive maintenance on the electrical grid. NOVA has three stages provided through a unified user interface: evaluation of input data quality, evaluation of machine learning and data mining results, and evaluation of the reliability improvement of the power grid. A prototype version of NOVA has been deployed for the power grid in New York City, and it is able to evaluate machine learning and data mining systems effectively and efficiently.

Categories and Subject Descriptors

D.2.4 [**Software**]: Software Engineering—*Software/Program Verification*; D.2.8 [**Software**]: Software Engineering—*Metrics*; H.2.8 [**Information Systems**]: Database Management—*Database Applications*

General Terms

Verification, Measurement

KDD4Service '11, San Diego, California, USA

Keywords

Data Quality Assurance; Performance Measurement; Machine Learning; Data Mining; Preventive Maintenance; Power Grid

1. INTRODUCTION

A sustainable energy future depends on an efficient, reliable and intelligent electricity distribution and transmission system, *i.e.*, power grid. The *smart grid* has been defined as an automated electric power system that monitors and controls grid activities, ensuring the two-way flow of electricity and information between power plants and consumers—and all points in between [6]. Without the smart grid, many emerging clean energy technologies such as electric vehicles and solar, wind or cogeneration power cannot be adopted on a large scale [2]. The smart grid of the future will have to operate efficiently to satisfy the increasing capacity demand, and should use the current legacy grid as much as possible to keep costs lower. This leads to a critical challenge of ensuring power grid reliability. In fact, the power grid has become less reliable and more outage-prone in the past years. According to two data sets, one from the U.S. Department of Energy and the other one from the North American Electric Reliability Corp., the number of power outages greater than 100 Megawatts or affecting more than 50,000 customers in the U.S. almost doubled every five years in the past fifteen years, resulting in about $49 billion outage costs per year [1]. A smart grid should anticipate and respond to system disturbances (self heal) proactively in order to minimize impacts on consumers.

To tackle this power grid reliability challenge, we have collaborated with the Consolidated Edison of New York, the main power utility provider of New York City, and developed several machine learning and data mining systems to rank some types of electrical components by their susceptibility to impending failure. The rankings can then be used for planning of fieldwork aimed at preventive maintenance, where the components should be proactively inspected and/or repaired in order of their estimated susceptibility to failure [14, 13, 8].

One important aspect of this type of collaborative research is that researchers and sponsors require objective evaluation of the machine learning and data mining model, the quality of the data input and output, and the consequential benefits, *i.e.*, physical system improvements, after the actions recommended by the machine learning and data mining have been taken. For this purpose, we have developed a comprehensive multi-stage online evaluation framework named *NOVA* (Neutral Online Visualization-aided Autonomic evaluation framework) that is able to provide such an evaluation objectively, effectively, and efficiently. We implemented NOVA in evaluating two complex online machine learning and data mining ranking systems for distribution feeders in New York City and analyzed the experimental results.

In the following section, we present preliminary information on the systems being evaluated. Then we describe the NOVA framework, followed by experimental results, analysis, and discussion. There is a large body of literature addressing machine learning and data mining algorithms for various domains, but little work describing how these algorithms should be evaluated in large complex systems; NOVA attempts to address this gap in the literature by providing an outline of how one might do this for the power grid. It can also be applied in other fields that have similar requirements.

2. BACKGROUND

2.1 Power Grid Failure

The power grid is the electricity distribution and transmission system that connects electricity generators and consumers. It is a power and information network consisting of power plants, transformers, high-voltage long-distance power transmission lines, substations, feeders, low-voltage local power lines, meters, and consumer appliances.

One of the main causes of power grid failure is electrical component failure. These component failures may lead to cascading failures. In 2004, the U.S.-Canada Power System Outage Task Force released their final report on the 2003 U.S. Northeast blackout, placing the main cause of the blackout on some strained high-voltage power lines in Ohio that later went out of service, which led to the cascading effect that ultimately forced the shutdown of more than 100 power plants [16].

2.2 Preventive Maintenance

To ensure the power grid is running smoothly, the electrical components that are most susceptible to failure should be proactively taken offline for maintenance or replacement. *Feeders* are transmission lines with radial circuit of intermediate voltage. In New York City, underground primary feeders are one of the most failure-prone types of electrical components. To predict feeder failures, we developed several machine learning and data mining systems to rank the feeders according to their susceptibility to failure.

MartaRank [3, 11] and ODDS [8] are two online machine learning and data mining-based feeder ranking systems for preventive maintenance. MartaRank employs Support Vector Machines (SVM), RankBoost, Martingale Boosting, and an ensemble-based wrapper. The ODDS ranking system uses ranked lists obtained from a linear SVM.

3. EVALUATION FRAMEWORK

NOVA conducts an automated and integrated evaluation at multiple stages along the workflow of an online machine learning and data mining system. There are three steps provided through a unified user interface, as illustrated in Figure 1: first, evaluation of the input data; second, evaluation of the machine learning and data mining output; third, evaluation of the system's performance improvement. The results from Step 1, 2 and 3 are eventually directed to a centralized software dashboard (a visualization-aided user interface). When abnormal results trigger pre-defined thresholds at any step, warning messages are dispatched automatically. We have implemented the NOVA evaluation framework for use on the New York City power grid, to conduct some comparative empirical studies on MartaRank and ODDS feeder ranking systems. In the following subsections, we will describe the details of each evaluation stage and demonstrate useful summarization charts for each step.

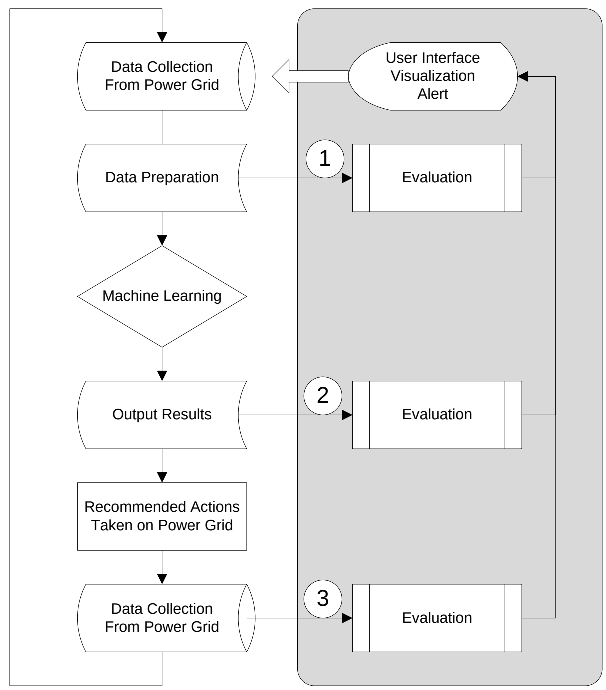

Figure 1: NOVA system design and workflow.

3.1 Step 1: Evaluation of Input Data Quality

In order for a machine learning and data mining system to perform as expected, the input data sets should meet predefined quality specifications. The evaluation process first uses *data constraints and checks* to see whether the required data exist and are up to date. Then the evaluation process conducts some more fine-grained checks, for example by using a *sparkline graph*, which is a type of information graphic characterized by its small size and high data density [15]. These checks would help researchers to correlate the changes in the input data sets with the variations of machine learning and data mining results, so that further study may be done to improve machine learning and data mining accuracy. As illustrated in Figure 2, for the one-day period preceding an actual outage, among ten feeder attributes—maximum scaled voltage, number of joints, number of cables, and peak load, etc.— being plotted, attribute 5 shows a big drop and subsequent climb in the sparkline time series graph. This kind of information can be important for feature derivation and selection.

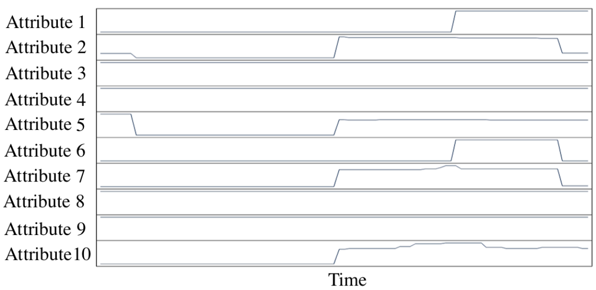

Figure 2: Sparkline graph for attributes data.

3.2 Step 2: Evaluation of Machine Learning and Data Mining Results

In evaluating a ranked list of components ordered by potential vulnerability, we use Receiver Operator Characteristic (ROC) curves, and accompanying rank statistics such as the Area Under the Curve (AUC). The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [4, 5]. It is in the range of [0, 1], where an AUC of 0.5 represents a random ordering, and an AUC of close to 1.0 represents better ranking with the positive examples at the top and the negative ones at the bottom. Figure 3 illustrates one typical ROC curve for a feeder ranking. The description for each data point highlighted on the curve (*e.g.*, 17M96 (511)) provides the feeder's name (*e.g.*, 17M96) and its ranking (*e.g.*, 511).

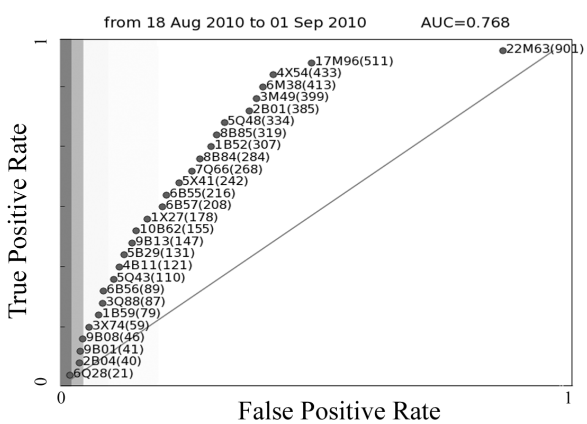

Figure 3: ROC Curve.

The ranking systems generate new models continuously, so the evaluation is presented as a time series of AUC values as shown in Figure 4. The black series in the figure shows the AUC time series of ODDS and the gray series shows the ones for MartaRank, both for the time period from May 2010 to November 2010. Our experiments show that MartaRank and ODDS feeder ranking systems have comparable overall performance according to the AUC.

3.2.1 AUC Cyclicity Challenge

One perplexing phenomenon we identified is the AUC cyclicity that appears in both feeder ranking systems as shown in Figure 4. Although the two AUC time series' vary differently, they both possess an inherent cyclical pattern, which we dubbed the *AUC cyclicity challenge*. It is an open problem to determine what causes this phenomenon. We hypothesize that an understanding of the mechanism behind the cyclicity challenge will lead to performance improvements for the systems.

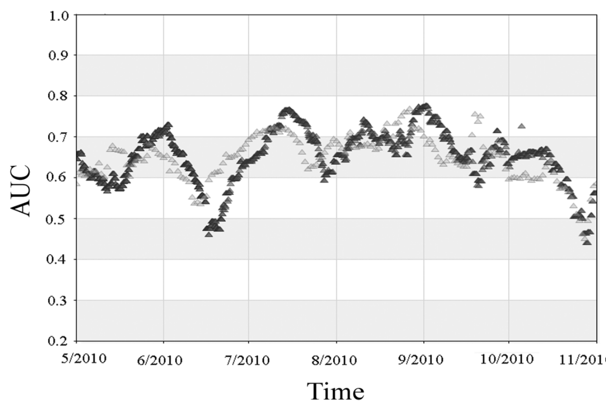

Figure 4: AUC time series graph.

3.3 Step 3: Evaluation of Reliability Improvement of the Power Grid

After the machine learning and data mining system outputs, the feeders ranked with highest susceptibility to failure are usually treated with a higher priority. The final stage of the evaluation is to validate that the recommended actions are in fact leading to the expected power system improvement, *i.e.*, fewer outages and longer time between failures. When considering longer time periods, a log(cumulative outages) versus log(time) chart is useful for seeing changes in the time interval between failures. This graphical analysis is also called a *Duane plot*, which is a log-log plot of the cumulative number of failures versus time [7], shown in Figure 5. The changing slopes of the regression lines show the improved rate of outages. If the failure rate had not changed, this log-log plot would show a straight line.

One experimental result we concluded from the evaluation using NOVA is the increasing MTBF (*Mean Time Between Failures*), *i.e.*, lower failure rate and better system reliability, for most networks. Figure 6 illustrates MTBF time series' for all of the feeders in a specific network for the period from 2002 to 2009 and the linear regression. Figure 7 illustrates the MTBF differences between 2009 and 2002 for each network. The bars with values above zero indicate MTBF improvements. The majority of the networks saw substantial increase in MTBF.

Table 1 lists the total number of feeder failures in the city from 2005 through 2009. The decreasing number of feeder failures means fewer outages for the power network.

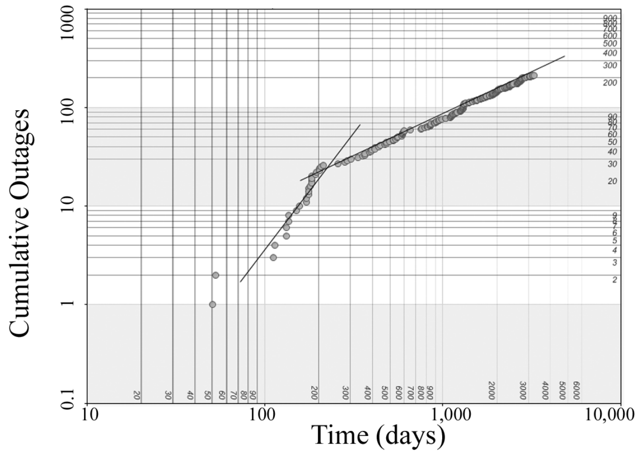

Figure 5: Cumulative outages versus time log-log chart.

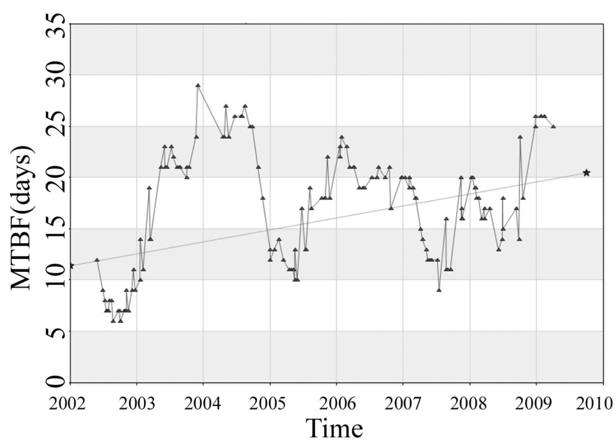

Figure 6: MTBF versus time and linear regression.

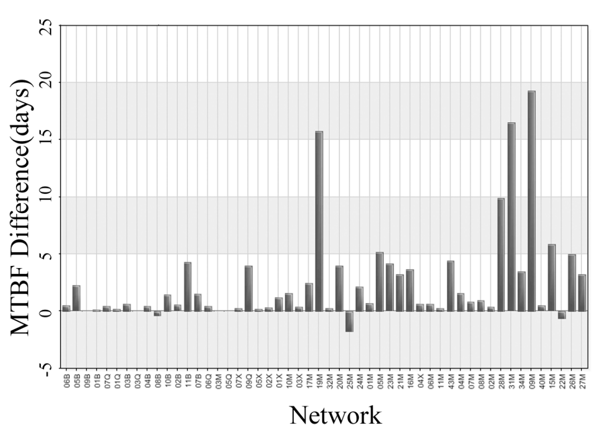

Figure 7: MTBF difference for each network.

Year	Number of Feeder Failures
2005	1612
2006	1547
2007	1431
2008	1239
2009	1009

Table 1: Number of feeder failures in the city.

Step	Evaluation target	Methods, metrics, charts
1	Input data	Sparkline graph, data checks and constraints
2	Machine learning and data mining results	ROC curve, AUC time series
3	Physical system improvements	Duane plot, MTBF, failure rate, linear regression
	Unified user interface	Dashboard, charts, triggers, warning messages, alert emails

Table 2: Summary of techniques used in evaluation.

To summarize the four key steps of the NOVA framework as described above, Table 2 lists the evaluation targets and main techniques (*e.g.*, methods, metrics, charts) used at each evaluation stage.

4. DISCUSSION

We have given examples above of each of the three steps of evaluation, using NYC power grid data. Depending on specific data and operational goals, there may be many ways to perform one of the three evaluations; the key point is that all of these three types of evaluation must be present. In machine learning and data mining, only the second type of evaluation is typically considered (step 2), and even that evaluation is mainly considered in static settings (without the element of time).

Langley's seminal paper "Machine Learning as an Experimental Science" made empirical study an indispensable aspect of machine learning research [10]. Since that time, many challenges in experimental machine learning have been identified. For instance, a more recent survey of Japkowicz reviewed shortcomings in current evaluation methods [9]. Through using NOVA on the New York City power grid, we have also been able to identify new challenges (*e.g.*, the AUC cyclicity challenge). In machine learning, the goal is often to optimize the criteria used for evaluation. NOVA suggests a much more ambitious set of evaluations than what are usually performed in machine learning and data mining experiments, potentially leading to a much broader way to consider and design machine learning systems, and hopefully leading to improvements in power grid operations.

Murphy *et al.* have studied verification of machine learning programs from a software testing perspective [12]. Our approach does not verify the internal correctness of the machine learning and data mining component. NOVA treats the machine learning and data mining process as a blackbox module and conducts its evaluations according to external specifications. This leaves the quality assurance of the machine learning and data mining software module to the machine learning researchers and software developers or testers.

5. CONCLUSION

Empirical evaluation of machine learning and data mining is important and challenging. This paper presented NOVA (Neutral Online Visualization-aided Autonomic evaluation framework), a framework that is able to evaluate real-time online machine learning and data mining applied in a complex mission-critical cyber-physical system objectively, effectively, and efficiently. The framework has been successfully experimented in evaluating machine learning and data mining for building a reliable power grid in New York City. Specifically, it was used to evaluate two complex feeder ranking systems and to generate predictions as the system was running. It has already proved to be a useful tool for machine learning and data mining researchers and smart grid control engineers. The NOVA framework can be applied to a wide variety of machine learning and data mining systems in which data quality, machine learning and data mining results, and reliability improvement are evaluated online.

6. ACKNOWLEDGMENTS

Wu and Kaiser are members of the Programming Systems Laboratory, funded in part by NSF CNS-0717544, CNS-0627473 and CNS-0426623, and NIH 2 U54 CA121852-06. Wu, Rudin, and Anderson (PI) are members of the energy research group at the Center for Computational Learning Systems at Columbia University. The group is funded by the Consolidated Edison Company of New York, U.S. Department of Energy and General Electric. We would also like to acknowledge funding from the MIT Energy Initiative.

7. REFERENCES

[1] S. M. Amin. U.S. electrical grid gets less reliable. *IEEE Spectrum*, page 80, January 2011.

[2] R. N. Anderson. Building the energy Internet. *Economist*, March 11, 2004.

[3] H. Becker and M. Arias. Real-time ranking with concept drift using expert advice. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 86–94, New York, NY, USA, 2007. ACM.

[4] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[6] Federal Smart Grid Task Force. Smart grid basics, 2010. Available at http://www.smartgrid.gov/basics.

[7] O. Gaudoin, B. Yang, and M. Xie. A simple goodness-of-fit test for the power-law process, based on the duane plot. *IEEE Transactions on Reliability*, 52(1):69–74, March 2003.

[8] P. Gross, A. Salleb-Aouissi, H. Dutta, and A. Boulanger. Ranking electrical feeders of the New York power grid. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pages 725–730, 2009.

[9] N. Japkowicz. Why question machine learning evaluation methods (an illustrative review of the shortcomings of current methods). In *2006 AAAI Workshop Evaluation Method for Machine Learning*. AAAI, 2006.

[10] P. Langley. Machine learning as an experimental science. *Machine Learning*, 3(1):5–8, 1988.

[11] P. M. Long and R. A. Servedio. Martingale boosting. In *Eighteenth Annual Conference on Computational Learning Theory (COLT)*, pages 79–94, 2005.

[12] C. Murphy, G. E. Kaiser, L. Hu, and L. Wu. Properties of machine learning applications for use in metamorphic testing. In *Software Engineering and Knowledge Engineering (SEKE)*, pages 867–872, 2008.

[13] C. Rudin, R. J. Passonneau, A. Radeva, H. Dutta, S. Ierome, and D. Isaac. A process for predicting manhole events in Manhattan. *Machine Learning*, 80(1):1–31, 2010.

[14] C. Rudin, D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, S. Ierome, D. Isaac, A. Kressner, R. J. Passonneau, A. Radeva, and L. Wu. Machine learning for the New York City power grid. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, May 2011.

[15] E. Tufte. *Beautiful Evidence*. Graphics Press, 2006.

[16] U.S.-Canada Power System Outage Task Force. Interim report: Causes of the August 14th blackout in the United States and Canada, 2003.