Managing Bad Data in Smart Electric Meter Measurements Using Data Mining Techniques

Glory Obielodan, Utah State University
Dr. Anurag Srivastava, Washington State University
Guo Yu, Washington State University
School of Electrical & Computer Engineering

Introduction:

It is not uncommon to find large error incidences in smart electric meter data collection. These errors can easily go unnoticed in large data sets and can lead to incorrect monitoring of electric power distribution system. Therefore, it necessary to detect bad data before the data is processed and possibly replace them with the best estimate.

Method:

In order to solve the above mentioned problem, basic statistical algorithms were utilized as described below. The following were implemented using Python, a programming language:

- First, I computed the average of the entire stream of data.
- Second, I used that average to find the standard deviation of the data.
- Third, I found a range of what would be considered “good” data (range = average ± standard deviation)
- I then scanned through the data, in sets of five values at a time and compared each value to the range of acceptable data.
- If any of the values fell out of the range they were either replaced with the average of the entire data stream or the average of two previous values, depending of their position in relation to the entire data stream.
- If the value was the first in the entire stream, it was replaced with the average of the whole data stream. If it was anywhere else, it was replaced with the average of the previous two values

Result:

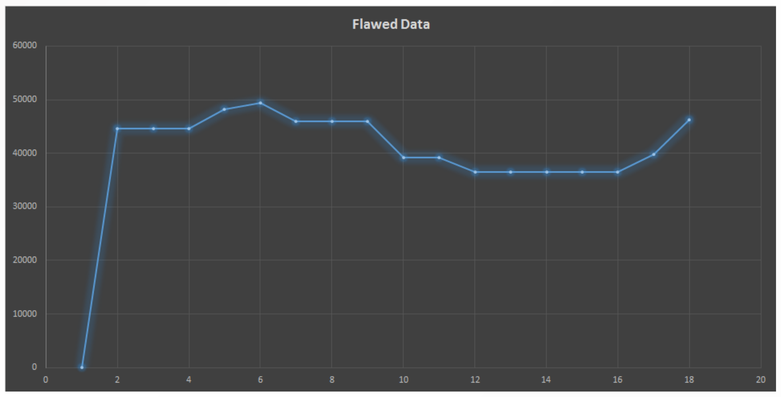

Chart 1: Flawed Data

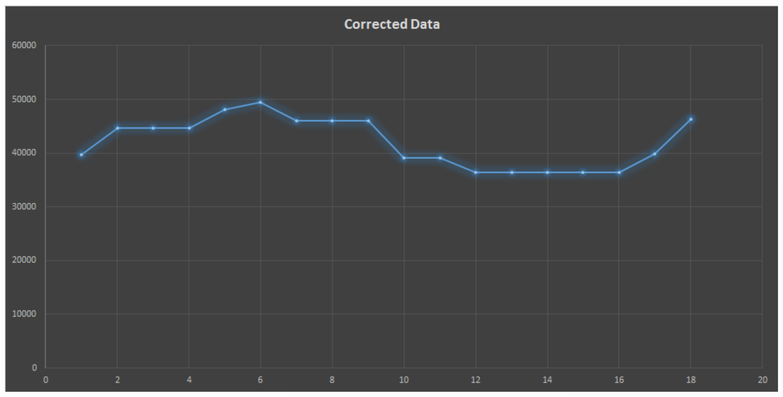

Chart 2: Corrected Data

The program seemed to work very well. It was tested using a small synthetic smart meter data, of two hundred and seventy numerical values with known errors, generated by electric distribution system analysis tool GridLab-D.
There was at least a 95% accuracy in its error detection and it replaced the errors with values that seem reasonable enough.

Displayed above are two charts that show the result of the program’s anomaly detection and correction. *Chart 1* is the flawed data, and *Chart 2* is the corrected version.

Conclusion/Discussion:

Overall, this project was a success. An algorithm was developed and implemented with a program that could detect large errors/anomalies in data streams using basic statistical methods. The program was able to detect said errors and correct them, with at least a 95% accuracy.

Further research has shown, however, that there exists far more advanced techniques that could increase the accuracy of the program’s anomaly detection and probably replace them with even better approximations. These methods utilize machine learning which is a branch of data mining that has a lot of good anomaly detection algorithms. In the future, further research on machine learning algorithms will be conducted.

Acknowledgements:

- Special thanks to Dr. Diane Cook and Chris Cain for their contribution to my research endeavors
- This material is based upon work supported by the National Science Foundation Research Experiences for Undergraduates Program under Grant No. 1460917.