As published in
UTILITYANALYTICS.COM March 2022

Using Artificial Intelligence to Support Utility Industry Solutions

By **Dominic Ciarlette,**
Machine Learning and Artificial Intelligence Focus Group Lead

EN DATA SOLUTIONS SEES ARTIFICIAL INTELLIGENCE AS A KEY TECHNOLOGY AS THE COMPANY CONTINUES TO INNOVATE SOLUTIONS FOR THE UTILITY INDUSTRY.

OVERVIEW

Data is the foundation of every decision-making process within the utility industry. Decision makers need high-quality data in order to make decisions that affect reliability, efficiency and safety. Utility asset data is particularly critical to make decisions involving the maintenance and operation of networks during normal operating and emergency conditions.

EN Data Solutions, a sector of ENTRUST Solutions Group sees artificial intelligence as a key technology as the company continues to innovate solutions for the utility industry. EN Data Solutions enhanced its already excellent data quality program by developing the Data Error Finder model, an advanced technology solution that improves data quality for utilities working with large asset datasets. The model improves data quality and review efficiency, which enables utility companies to strengthen their data foundations.

CONTEXT

More data does not necessarily mean more information. Data is prone to errors, particularly in large and complex asset datasets. Types of data errors include incorrect asset attributes, contradicting attributes, missing information, duplication, formatting violations, etc. The causes behind data errors often can be attributed to human error during manual data entry. Other causes include varying systems with different importation protocols, data standards changing over the life of the dataset, transcription errors in automated data entry (when a computer extracts information from scanned documents), out-of-date datasets, file corruptions and errors introduced through data mismanagement over time.

Data errors have consequences. The effects of poor data quality permeate every layer of an operation and, because utilities deal with high consequence situations, it is vital that decisions rest on a solid data foundation.

The mix of numerical and categorical data, typically found in the utility asset data, introduces challenges to finding and correcting errors. That is why there is a need for a mechanism that can be used to sift data for suspicious records in order to identify mistakes, and for quality checks to be implemented so that future operations prevent them from reoccurring.

It can be a painstaking process to manually sift through suspicious data in search of errors. In a large dataset, 90 percent to 95 percent of the data often is accurate. Correcting the errors is not the difficult part; finding the errors presents a much greater challenge.

SOLUTION

Against this backdrop, EN Data Solutions investigated how machine learning can support cleansing data efforts. While testing proven error-detection methods, EN Data Solutions found that leveraging outliers and data self-contradiction are two approaches, which were well suited for application to utility asset data. However, utility asset data often is categorical in nature (material type/grade, how it was welded, etc.), rather than numerical. This increases the complexity of comparing like-for-like records. Fortunately, machine-learning clustering techniques work well with categorical data, where traditional techniques fall short.

EN Data Solutions developed the Data Error Finder to mimic subject matter experts' familiarity with datasets. The company does not supply the Data Error Finder with an explicit rule set to follow, such as where specific assets can be and what attributes they can have. Instead, the Data Error Finder learns that knowledge itself by using the entire dataset.

The Data Error Finder accomplishes this learning by applying a cluster algorithm to the utility asset records. The results of this clustering reveal rule-breaking outliers hidden within the dataset. This information can be used to perform a targeted quality review and update data rule sets. It also can be used to supplement existing data quality control efforts.

CLUSTERING

Clustering is an unsupervised machine learning method that allows EN Data Solutions to take high-dimensional data and evaluate internal relationships. The dataset can be presented to the model, and the records can be grouped together based on similarity. Fortunately for the utility industry, this method can be applied to both categorical and/or numerical data. Clustering serves as a time-saving technique that is compatible with datasets on a scale of thousands to millions of records. It also can accommodate more than 80 columns of tabular or encoded spatial information. The aim is to look for records in the dataset that have a high probability of containing errors.

The Data Error Finder returns a list of outlying records per cluster (Figure 1). Records lying significantly outside of its cluster have a high probability of containing errors and these records should be targeted for further investigation. This method can be used to assist with data quality checks during and after creating or updating data. It also can be used for evaluating the current health of an existing dataset, by identifying records that are outliers from the cluster.

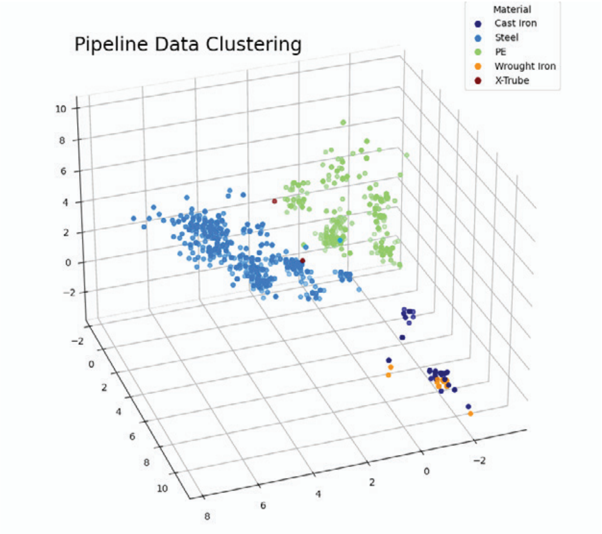

Figure 1. Each point above represents a record. The spatial relationship between points implies similarity.

IMPACTS

Since manual data review is extremely time-intensive and error-prone, the ability to identify records for closer inspection is invaluable. EN Data Solutions is currently engaged in a project in which the Data Error Finder will save approximately 500 hours of manual data error review. This approach will benefit any company that has large datasets that were created using manual data entry methods or that contain many similar records.

Overall results and time saved will vary depending on the health of the dataset the system is working on. If, for example, a high-quality dataset is 99% perfect, the Data Error Finder may reveal the elusive 1% that contains errors. In these cases, it can be difficult for a utility to justify performing a thorough review of every record for an already 'healthy' dataset. The Data Error Finder lowers the cost of finding errors in a relatively healthy dataset by identifying those records most likely to contain errors.

EXAMPLES

A client approached EN Data Solutions with an expansive dataset covering more than 25,000 miles of pipe to review for quality. After running the Data Error Finder, five main clusters were identified (Figure 2).

Once the clusters were established, the outliers could be identified and examined. For example, in Cluster 3, some obvious contradictions in attribute values were found in the material, grade and joint type fields in the records most distant from the middle of the cluster (Figure 3).

These records clearly contained errors. Once these errors were identified, a rapid check for similar errors in the data was performed. By targeting known error types, the amount of time needed to review the entire dataset was substantially reduced. Significantly, these errors were found in 6% of the records. Rather than review the entire dataset for errors, the client was able to focus corrective action on the relatively small number of records that contained errors.

In another recent project, EN Data Solutions was asked to review a client’s service line records because they had encountered many errors in the dataset. At the end of the project, after EN Data Solutions had conducted a manual error analysis on a subset of the dataset, the client learned that approximately 15% of their records did not contain any errors. Extending this error rate to the entire dataset of 60,000 records suggested that 9,000 would probably not require corrective actions. Applying the Data Error Finder to the remaining 60,000 un-reviewed records could reduce the effort necessary to separate the correct and erroneous records. Such a result could enable a focused application of effort and a significant time saving.

TECHNICAL DETAILS

The basic machine learning solution approach used for the Data Error Finder involved first identifying a problem to solve, then gathering data, applying feature engineering, choosing an algorithm, training, validating, revising and applying the solution.

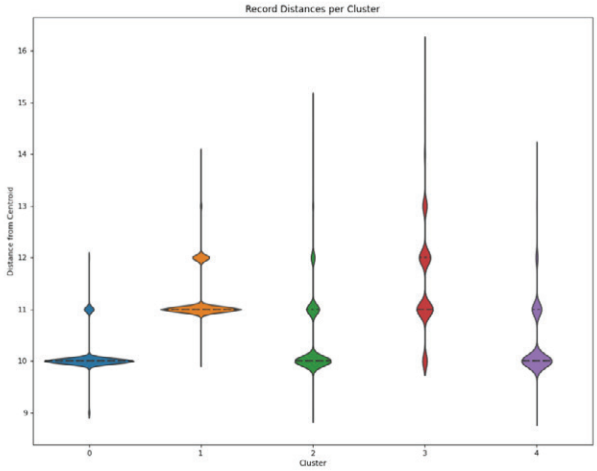

Figure 2. The widths of the above figures indicate the magnitude of records at a given distance from each cluster.

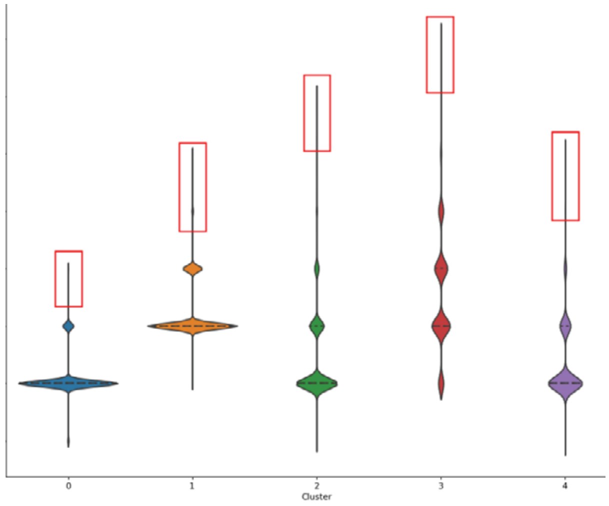

Figure 3. The outliers are found at the narrow tops of these figures.

The objective of the feature engineering step is to transform data into a language that the computer can understand, while preserving the information contained in the data. Simple techniques of feature engineering include standardization and using data manipulation tools to cleanse the data. A challenge unique to the utility space is encoding and preserving the information present in the spatial aspect of asset datasets. EN Data Solutions utilized its geographic information system (GIS) expertise to meet the challenges of spatial information encoding. This greatly expanded the available dimensions of the dataset and provided the model with more training data.

Following feature engineering, EN Data Solutions trained and validated the clustering-based model by using the target dataset. In general, data clustering involves the measurement of similarity between records to group or cluster records together. EN Data Solutions used these similarity measurements as the basis of the solution outputs.

FUTURE DEVELOPMENTS: COMPUTER VISION APPLICATIONS

EN Data Solution’s newest development is leveraging computer vision techniques to support spatial analysis solutions. Computer vision involves the training of a model to identify objects or areas within imagery data and replicating or improving upon the abilities of a person given the same task. In the utility space, this typically involves aerial and satellite imagery. Such computer vision techniques may become a powerful tool for initiatives such as operations support, asset maintenance, risk assessments, load forecasting and service area categorization. EN Data Solutions is working to apply computer vision solutions to its existing and future service offerings in order to take advantage of the potential efficiency, accuracy and reliability improvements presented by these technologies (Figure 4).

SUMMARY

EN Data Solutions benefits from a unique position at the intersection of many aspects of the utility industry. By taking advantage of its range of industry expertise, EN Data Solutions stays abreast of emerging technologies and the application of those technologies to its clients’ problems. The Data Error Finder is one example of an advanced data solution that drew upon several skills, including data domain expertise, geospatial proficiency and advanced software development.

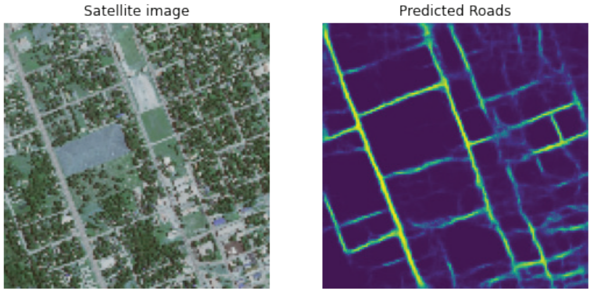

Figure 4. Example of road identification using computer vision.

The Data Error Finder uses machine learning to cluster utility asset records by data similarity, revealing rule-breaking outliers that are hidden within the dataset. This information can be used to perform a targeted quality review and update data rule sets, as well as supplement existing data quality control efforts. The Data Error Finder has a geospatial asset focus but is applicable to many large datasets that feature a mix of categorical and numerical data types, which includes almost all utility asset record datasets. However, the tool can be used in any industry that draws upon a large amount of data that contains errors.