Machine Learning for Automated Metadata Assignment in Buildings

Cooperative Research and Development Final Report

CRADA Number: CRD-18-00767

NREL Technical Contact: Dylan Cutler

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-7A40-82451
March 2022

Machine Learning for Automated Metadata Assignment in Buildings

Cooperative Research and Development Final Report

CRADA Number: CRD-18-00767

NREL Technical Contact: Dylan Cutler

Suggested Citation
Cutler, Dylan. 2022. *Machine Learning for Automated Metadata Assignment in Buildings: Cooperative Research and Development Final Report, CRADA Number CRD-18-00767.* Golden, CO: National Renewable Energy Laboratory. NREL/TP-7A40-82451. https://www.nrel.gov/docs/fy22osti/82451.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-7A40-82451
March 2022

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Cooperative Research and Development Final Report

Report Date: March 15, 2022

In accordance with requirements set forth in the terms of the CRADA agreement, this document is the CRADA final report, including a list of subject inventions, to be forwarded to the DOE Office of Scientific and Technical Information as part of the commitment to the public to demonstrate results of federally funded research.

Parties to the Agreement: RealTerm Energy U.S. LP

CRADA Number: CRD-18-00767

CRADA Title: Machine Learning for Automated Metadata Assignment in Buildings

Responsible Technical Contact at Alliance/National Renewable Energy Laboratory (NREL):

Dylan Cutler | Dylan.Cutler@nrel.gov

Name and Email Address of POC at Company:

Sean Neely | s.neely@brainboxai.com

(for Jean-Simon Venne | js.venne@brainboxai.com)

Sponsoring DOE Program Office(s): Office of Energy Efficiency and Renewable Energy (EERE), Building Technologies Office (BTO)

Joint Work Statement Funding Table showing DOE commitment:

No NREL Shared Resources

Estimated Costs	NREL Shared Resources a/k/a Government In-Kind
Year 1	$.00
TOTALS	$.00

Executive Summary of CRADA Work:

RealTerm Energy and NREL have identified a shared vision to evaluate opportunities to facilitate the organization and assignment of metadata to building control system (BCS) data via industry-informed machine learning (ML). Manual metadata assignment is labor intensive and costly, slowing down any Energy Management and Information System (EMIS) deployment in the building space. This project aims to develop methodologies to accurately assign this metadata and significantly decrease the level of effort associated with deploying EMIS.

The objective of this project is to identify/design methodologies to assign metadata to HVAC control points automatically. The identified methodologies will be programmed in analytics algorithms so they can ingest a list of points and produce a detailed tagging following the Haystack[1] classification nomenclature. To validate the efficacy of each methodology, tagging results will be compared utilizing a list of points extracted from RealTerm's building database—as well as data extracted from the NREL campus via the Intelligent Campus program—enabling testing against large datasets with real world challenges. The developed methodologies may leverage building manager/operator input on a limited basis to add context to the classifying algorithms.

The partnership aims to advance global efforts in areas related to the DOE missions through improving operational performance of commercial buildings. It is well documented that buildings fall out of commission after they are occupied, wasting significant energy and incurring associated costs simply due to poor operational performance. Emerging EMIS technologies that perform continuous commissioning help to address this issue, yet integration of these systems can be labor intensive both for the technology vendor and the building owner/operator. This project will enable more efficient and cost-effective analytics for buildings, enabling improvement in building operations at lower cost points.

Summary of Research Results:

Task 1: Identify/design the different methodologies to be tested

This task combined a literature review of the work in automated metadata tagging with a detailed review of RealTerm Energy's requirements for integration with their existing software stack and approaches. We then moved into methodology selection and design based on that review and on specific requirements of the tagging process. There were two main categories of data addressed in the literature: (1) BCS point name data and (2) the time series data associated with a given point.

Most of the prior work in BCS point name analysis had focused on rule-based evaluation of names, combined with user-defined dictionaries. This had led to significant human intervention, required the manual dictionary development, and did not accommodate multiple languages. To address these concerns, we identified a text-based grouping and clustering approach called "k-mers". This was taken from DNA encoding research and works by generating all combinations of k-mers (essentially three or more character groupings within the larger text string) and finding similarity indices between different point names to group them into clusters.

[1] https://project-haystack.org/

To apply tags based on timeseries data from a BCS point, most prior work had leveraged traditional ML approaches (both supervised and unsupervised) in clustering and classification for tag application. Two primary steps were identified for applying these ML algorithms: (1) feature extraction and (2) ML algorithm implementation. Feature extraction consists of calculating statistics regarding the timeseries data that can be put into feature vectors upon which the ML algorithms can be trained and then run. The ML algorithms generally fall into supervised or unsupervised categories with supervised algorithms being used more often with the times series data. During this stage of the project, we identified an initial set of features that we would extract from the data and a suite of ML algorithms that we would implement, focusing on supervised, classification algorithms (e.g., random forest and support vector machines).

Additionally, at this stage in the project we decided to focus our efforts in applying Project Haystack tags to the points themselves, but not addressing the additional challenge of grouping points into pieces of equipment. This will be addressed in future work.

Task 2: Program select methodologies

The literature review and algorithm identification step was followed by implementation in a coding environment. We choose to implement in a python environment both for compatibility with the code base used by RealTerm Energy U.S. team, and to leverage the large ML packages that python supports (skikit-learn was leveraged in this project).

First, we implemented a set of classification ML algorithms including random forest, logistic regression, and support vector machine approaches for classifying the times series data. To test these algorithms, we needed to implement the feature extraction component. Initially, we extracted mean, median, minimum, maximum, variance on an hourly frequency and combined those into feature vectors. We then extended that to test additional features, including testing those features over varying time windows (1-hr, 4-hr, 1-day and found that 1-hr performed best), adding additional variables (e.g., skewness, kurtosis, interquartile range, and derivative), testing overlapping windows (50% overlap in with two-level statistical abstractions), and other shape-based or model-based methods (e.g., empirical mode decompositions). The implemented ML algorithms were then tested against many different combinations of the feature vectors for optimizing feature and algorithm selection.

Additionally, we implemented the unsupervised, k-mers clustering methods for grouping point names into like clusters. As part of this implementation, we included a new measure of k-mer similarity that better preserves ordering of the k-mers in the larger word that has been decomposed. The vectorized strings are then clustered using an agglomerative hierarchical clustering approach. Figure 1 shows the results of this k-mers clustering approach for all of the points in a single building. The similarity metric and associated clustering methods are described in more detail in the journal article published as part of this CRADA [1].

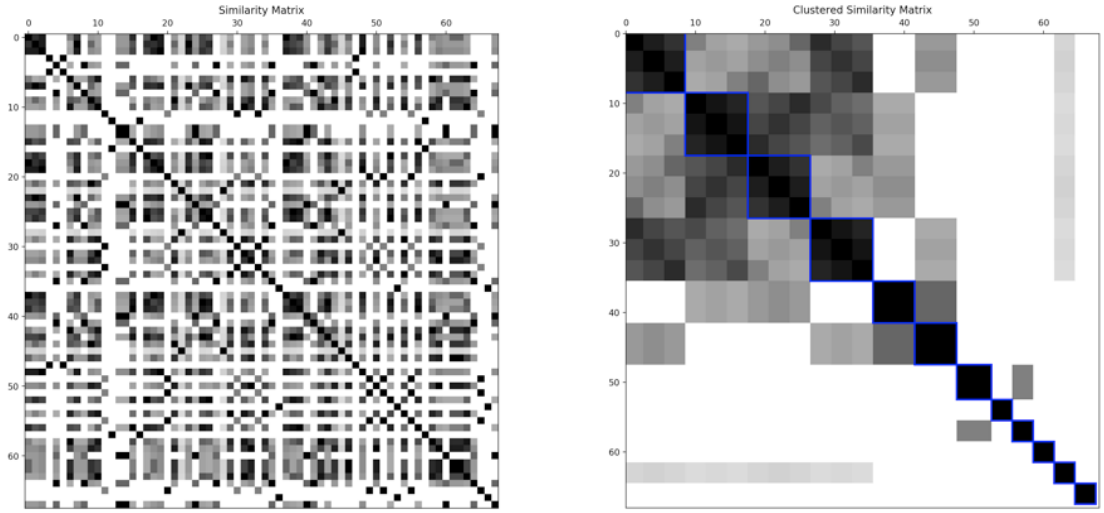

Figure 1. Unclustered and Clustered similarity matrices for the point name k-mers

Task 3: Define testing data sets and evaluate algorithm performance

To effectively test the performance of the algorithms implemented in Task 2, we compiled two different sets of BCS points and associated timeseries data. The first dataset consisted of three different commercial retail buildings from RealTerm Energy's building portfolio. Each of the buildings had 2-3 roof top units along with a number of interior space sensors, totaling approximately 40-60 points per building. The second dataset was taken from the Energy Systems Integration Facility (ESIF) building on the NREL campus. This dataset consisted primarily of the larger air-handling systems in the ESIF and contained 352 unique points. Each of these sets of BCS points had to be authoritatively tagged according to the Project Haystack metadata standard. This was performed by the project team and enabled execution of the supervised ML algorithms and evaluation of the different algorithms' performance.

We tested the different algorithms, combined with different combinations of extracted feature sets. We assessed performance of the algorithms based on the true positives, false positives, and false negatives generated through tag application by the algorithm as well as by the F1 scores. We include the F1 score results for two of the supervised ML algorithms (random forest and support vector machine) as compared to the unified architecture approach (described below in Task 4) in Table 1.

Task 4: Evaluate ensemble algorithms

Application of metadata in buildings is a challenging problem that is fraught with challenges introduced by varying BCS implementations and associated software, inconsistent point naming approaches, complex building systems, and human errors (e.g., misnamed points, incorrect units, etc.). Therefore we did not expect to sufficiently address this problem through a single ML algorithm or approach, instead we expected to utilize multiple methods and incorporate them into an "ensemble" algorithm. This ensemble algorithm would likely still require human intervention to finalize the tagging or deal with especially unique or complex systems, but would reduce the effort required significantly. This task was included to explicitly evaluate opportunities in this space.

To address these challenges we ended up developing what we called a "unified architecture" (UA) that incorporated most of the individual methodologies described in task 2. Additionally, it incorporated rule-based logic to (1) determine what sets of tags were candidate tags for the building in question (and if there are sets of mutually exclusive tags, such as sensor, cmd, sp), and (2) how to combine the results from the different supervised algorithms—in particular random forest (applying all tags at once) and supervised classifiers such as support vector machines (applying a single tag at a time)—such that their combined results are utilized to predict tag application. Additionally, the k-mers clustering approach is integrated into the UA to provide a level of confidence on the tags that were applied and flag ones that should potentially be revisited. The complete UA flow diagram is presented in Figure 2.

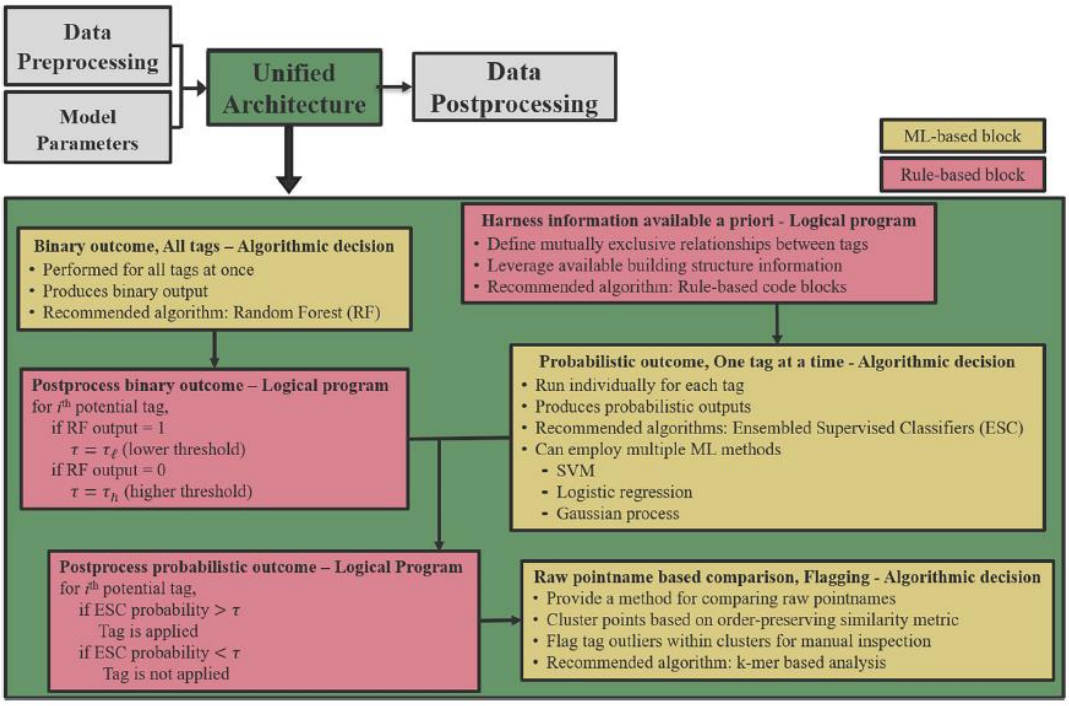

Figure 2. Block diagram of the UA to highlight various components

The UA was programmed into an end-to-end python workflow and resulted in a software record (Unified Architecture for Automated Building Metadata Tagging using Machine Learning" NREL Software Record SWR-20-38).

Task 5: Analyze the results and compare methodologies

The final task in the CRADA was to analyze the results generated by the project, specifically the ensemble approach that we codified into the UA workflow. We tested the individual algorithms and the UA approach against the three different commercial retail buildings as well as the ESIF dataset and assessed algorithm performance. The results for the three commercial retail buildings are summarized in in Table 1. The algorithms did not perform quite as well with the more complex ESIF dataset but the UA was still able to correctly apply 70-75% of the tags and obtained an F1 score of >0.8 in all test cases.

We can note from Table 1 that the ability to combine the different ML algorithms, incorporate rule-based logic, and leverage unsupervised k-mers clustering improves the F1 scores for all buildings relative to the individual algorithm approach.

Table 1. F1 Scores for tag application by individual methods versus the unified architecture developed under this CRADA for three different buildings

Approach	Bldg. A	Bldg. B	Bldg. C
Random Forest	0.71	0.71	0.69
Support Vector Machine	0.83	0.84	0.85
Unified Architecture	0.90	0.90	0.88

Future work will evaluate the potential to incorporate "template" type approaches that are being developed in projects such as Haste [2] and will address the challenges associated with identifying relationships between points. This will require identifying and tagging pieces of equipment, and grouping points onto those pieces of equipment, essentially identifying and applying the "relationship" tags that group points into pieces of equipment.

References

[1] S. Mishra, A. Glaws, D. Cutler, S. Frank, M. Azam, F. Mohammadi, J.S. Venne (2020) Unified architecture for data-driven metadata tagging of building automation systems. *Automation in Construction* **120** (1-14)

https://www.osti.gov/pages/biblio/1669624

[2] C. Mosiman and A. Viveiros (2020) Haste Software Code. U.S. DOE Office of Energy Efficiency and Renewable Energy (EERE), Energy Efficiency Office. Building Technologies Office https://www.osti.gov/doecode/biblio/41387

Subject Inventions Listing:

None

ROI #:

None