Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230722

Commercial HVAC system fault diagnosis using big data analytics: A case study

Yuhang Liu[0000-0002-7202-0700]

University of California-Riverside, Riverside, CA 92521, USA

yliu800@ucr.edu

Abstract. With the increasing use of heating, ventilating, and air conditioning (HVAC) systems nowadays, their energy consumption is receiving more attention. The study begins by trying available anomaly detection techniques, including KNN, COF, and isolated forests. The comparison reveals that these methods disregard some linear correlation results. Then, the pattern is summarized by analyzing data from 100 HVAC-equipped rooms. Next, the study uses correlation analysis and neural networks to identify abnormal HVAC data. Finally, it concludes by analyzing the factors that lead to the anomalies.

Keywords: HVAC system anomaly detection, correlation analysis, fault diagnosis, data analysis.

1. Introduction

Due to the complexity of modern heating, ventilating, and air conditioning (HVAC) systems, approximately 40% of buildings have misconfigured equipment, wasting up to 40% of energy [1]. The buildings also have very large and complex systems that contain thousands of moving parts, such as dampers, water pumps, and fans [2]. Therefore, if such a complex system has problems, the loss brought about is also huge. The complexity of modern heating, ventilating, and air conditioning (HVAC) systems results in misconfigured equipment in about 40% of buildings, wasting up to 40% of energy consumption [3]. Even if improvements in sensing, communication and control infrastructure will improve operational efficiency [4], this still increases the complexity of the system, which may lead to increased energy loss. In addition to energy waste, HVAC failures can lead to occupant discomfort, reduced indoor air quality, and depreciation of equipment [5]. These errors and wastes can be considered as exceptions. Therefore, if these faults are detected in time using anomaly detection, up to 40% of energy can be saved and useless loss can be reduced [3].

There are a variety of anomaly detection techniques that can be used, including KNN, COF, and Isolation Forest [6–8], and each of these techniques has its own benefits for spotting anomalies in the temporal data of HVAC systems. However, because they do not offer a thorough analysis of the relationship between the various columns, these techniques run the risk of ignoring some relevant information [9]. This paper applies correlation analysis to HVAC systems in order to address this issue. The results are then used to train neural networks, and model predictions are then used to assess the analysis's findings.

The studies discussed in the paper consist of 5 sections. Section 2 describes the background related to HVAC systems. Section 3 shows the methods this paper used and the laws summarized. Section 4

DOI: 10.54254/2755-2721/13/20230722

uses the result of the correlation analysis to train neural networks and check the prediction result, and Section 5 makes the conclusion.

2. Background Research

2.1. Industry overview

In 2019, there were approximately 123 million buildings across the United States. More than 80% of them are 20 years or older. They are responsible for the consumption of 40% of the nation's total energy supply and 75% of its electrical production. At least 20% of this energy is wasted as a result of inefficiency in the design and operation of the building [10]. It was estimated that HVAC systems consumed 57% of the total energy used in commercial and residential buildings in the United States in 2011[11]. The HVAC system is still not efficient enough to justify the amount of energy it uses. Abnormalities in HVAC systems can result in a significant amount of wasted energy, which is a significant cause of suboptimal performance. Due to the fact that each HVAC system has unique patterns and produces unique results, the majority of existing algorithm-based methods for detecting anomalies cannot be directly applied to HVAC systems. With the goal of making the model more applicable to HVAC systems as a whole, this paper takes a data-driven approach to enhancing the model's capabilities for anomaly detection.

2.2. Existing approaches

Issues with a HVAC system are considered abnormalities. Anomaly detection refers to the processes used to identify outliers. Anomaly detection can be defined in many ways; one definition is "an outlying observation is one that appears to deviate markedly from other members of the sample in which it occurs" (Grubbs [1]). Anomaly detection, if used effectively, could help HVAC systems catch problems before they cause excessive energy consumption.

The following are some examples of methods used to spot anomalies:

2.2.1. KNN (K-Nearest Neighbors). The basic idea behind this method is to find the k training samples that are closest to the new sample by calculating the distance between the two samples [12]. However, a suitable value for k must be chosen, and the success of the classification is highly dependent on this value.

2.2.2. Connectivity-Based Outlier Factor (COF). The local density of COF is calculated based on the average chain distance. It begins by calculating the k-nearest neighbor of each point, followed by the Set-based nearest Path of each point [12–13]. The results are then used to calculate the chain distance, followed by the COF value, which can be used to determine the proportion of outliers and the number of nearest neighbors.

2.2.3. Isolation Forest. The Isolation Forest algorithm divides a data space into two subspaces using a random hyperplane and a single cut. The algorithm then continues to randomly select hyperplanes to divide the two subspaces obtained in the first step, until each subspace contains a single data point. It divides the plane to further isolate the isolated points, then creates an isolated tree, evaluates the data using the isolated tree, and calculates the anomaly score using an algorithm. The central premise of its calculation is that anomalous samples fall into leaf nodes more quickly and easily [14].

Consequently, this paper endeavors to analyze the data using simple correlation analysis techniques, to draw conclusions from the data results, and to investigate the causes of the anomalies.

DOI: 10.54254/2755-2721/13/20230722

3. Methodology

3.1. Dataset overview

In this case, data was collected from a commercial healthcare facility in the United States. The dataset contains data for the HVAC systems in a total of 100 rooms, including the following information: Data Time, Airflow, Airflow Setpoint, Damper Position, Discharge Air Temperature, Zone Temperature, Hot Water Valve Command, and Flag.

3.2. Data analysis

The primary focus of the data analysis was on data relationships with moderate or higher correlation. The findings were summarized as follows: damper, airflow, and airflow setpoint have positive correlations that range from moderate to strong; hot water command and discharge air temperature have positive correlations that range from moderate to strong. The values of the Pearson coefficients between each characteristic that were derived from the correlation analysis are presented in Table 1 (the median value of the training rooms).

Table 1. The Pearson coefficients between each characteristic.

	AF	**AFS**	**DP**	**DAT**	**ZT**	**HWVC**
AF	1	0.506	0.703	-0.083	-0.054	0.078
AFS		1	0.764	-0.037	0.010	0.082
DP			1	-0.056	-0.021	0.062
DAT				1	0.171	0.715
ZT					1	0.009
HWVC						1

AF: Airflow; **AFS:** Airflow Setpoint; **DP:** Damper Position; **DAT:** Discharge Air Temperature **ZT:** Zone Temperature; **HWVC:** Hot Water Valve Command

To determine the relevance standards, an approach based on quadratic equations was used, and the following are the anomaly standards that were derived from the analysis:

Table 2. Anomaly Standards Based on Quadratic Equations

airflow - damper position	0.490
airflow - airflow setpoint	0.275
damper position - airflow setpoint	0.621
hot water command - discharge air temperature	0.467

It is generally accepted that environmental factors such as room size, initial room temperature, and ventilation are to blame for the lack of a significant linear correlation between the temperature of the discharged air and the temperature of the zone.

In the course of the research, it was discovered that despite the fact that Airflow and Damper Position have a correlation coefficient that is, on average, 0.764, there are still some regions in which the correlation coefficient is significantly lower than 0.764. There are a number of other data sets that exhibit a situation that is very similar to this one. When looking at the raw data, it is discovered that the majority of them are attributable to the following reasons:

3.2.1. Data time lag. Because there is a lag between the data, a high correlation will be found if a column is panned by several minutes after another column.

3.2.2. Error in data collection. There is no discernible movement at all in the data, which may indicate that there is an issue with the configuration of the system or that there is an error in the sensitivity of the sensor.

DOI: 10.54254/2755-2721/13/20230722

3.2.3. Bias in Person coefficients. Since this case study employs primarily linear correlation Pearson coefficients, the results will be biased if there is a strong nonlinear correlation.

3.3. Approach to data processing

In this paper, correlation analysis is the main method for processing data. Calculating the correlations in the various parts, such as between airflow and damper position of each room, performing box plot analysis on these parts, and deriving a standard line for coefficients outside the normal interval of the box plot overall. This analysis employs the Pearson correlation coefficient. Finding the quartiles of the dataset is the primary method for locating anomalies in boxplots. The outliers are identified by grouping the data by hour and then adding a 'flag' column by comparing the correlation of each group with the overall correlation (1 for abnormal and 0 for normal).

As a result of graph analysis of the data for each room, the following characteristics are determined:

3.3.1. Baseline in Airflow. Airflow is constantly changing, but based on a baseline, and the change is approximately 10%.

3.3.2. Baseline in Discharge Air Temperature. Analysis reveals that Discharge Air Temperature fluctuates around a baseline, which varies by approximately 30%.

3.4. Neural networks

Two varieties of neural networks are used for training here:

3.4.1. Simple neural network model from Keras Sequential model package. Utilize this local model for the binary classification problem with the 'rmsprop' optimizer.

3.4.2. The MLPClassifier in the Sklearn package. A supervised learning algorithm, feed-forward artificial neural network model, which is, in essence, a fully connected neural network. This algorithm is implemented effectively within the Sklearn package.

4. Validation

I selected 90 out of 100 training rooms and 10 testing rooms for validation purposes. Training with these two neural networks demonstrates that there is little difference in the prediction accuracy of the trained models, proving that the models are capable of convergence. Following a comparison of the accuracy of the two neural networks mentioned previously, the accuracy is displayed below:

Table 3. Accuracies of two neural networks.

	Sequential Model	MLPClassifier
Airflow - Damper Position	80.93%	81.24%
Airflow - Airflow Setpoint	88.73%	90.19%
Damper Position - Airflow Setpoint	66.55%	71.97%
Discharge Air Temperature - Hot Water Valve Command	79.42%	79.69%

5. Conclusion

This article is a case study of research into HVAC (heating, ventilation, and air conditioning) anomaly detection. The paper begins by providing a summary of the patterns of anomalies by conducting correlation analysis, which is the study of the correlation between columns in the system data. Next, the paper summarizes those columns that produce anomalies by looking at the content of the raw data and then analyzing it using graphs. After importing the patterns into the neural network to train a model, the results are then compared to previous versions of the model.

DOI: 10.54254/2755-2721/13/20230722

This case study performs only an analysis of linear correlations in order to search for anomalies; it does not take into full consideration an analysis of non-linear correlations. In addition, the time delay of the data was not taken into consideration when the data were being handled. The next step could involve more in-depth data processing, including time delay and three or more dimension feature correlation analysis. In addition to that, we could try out additional neural network models, such as regression models.

References

[1] F. E. Grubbs, “Procedures for detecting outlying observations in samples,” Technometrics, vol. 11, no. 1, pp. 1-21, February 1969.
[2] ASHARE. Handbook of HVAC system and equipment. Technical report, 1996.
[3] B. Narayanaswamy, B. Balaji, R. Gupta and Y. Agarwal, “Data driven investigation of faults in HVAC systems with model, cluster and compare(MCC)” BuildSys@SenSys, page 50-59. ACM, (2014).
[4] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, “Stl: A seasonal-trend decomposition procedure based on loess,” Journal of Official Statistics, vol. 6, no. 1, pp. 3–73, 1990.
[5] E. Mills. Building commissioning: A golden opportunity for reducing energy costs and greenhouse-gas emissions. Lawrence Berkeley National Laboratory, 2010.
[6] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in ACM SIGMOD Record, vol. 29, no. 2. ACM, 2000, pp. 427–438.
[7] M. Goldstein and A. Dengel, “Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm,” KI-2012: Poster and Demo Track, pp. 59–63, 2012.
[8] M. Leng, X. Chen, and L. Li, “Variable length methods for detecting anomaly patterns in time series,” in Computational Intelligence and Design, 2008. ISCID 08. International Symposium on, vol. 2. IEEE, 2008, pp. 52–56.
[9] M. Munir, S. Erkel, A. Dengel and S. Ahmed, "Pattern-Based Contextual Anomaly Detection in HVAC Systems," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 1066-1073, doi: 10.1109/ICDMW.2017.150.
[10] DOE Building Technologies Office Overview, https://www.gsa.gov/cdnstatic/Bouza%20-%209-12-19%20BTO%20overview.pdf, last accessed 10/15/2022.
[11] Siyu Wu, Jian-Qiao Sun, “Cross-level fault detection and diagnosis of building HVAC systems,” Building and Environment, Volume 46, Issue 8, 2011, Pages 1558-1566, ISSN 0360-1323.
[12] Cunningham, Pádraig and Sarah Jane Delany. “k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples).” ArXiv abs/2004.04523 (2020): n. pag.
[13] Nowak-Brzezińska, A., & Horyń, C. (2020). “Outliers in rules - the comparison of LOF, COF and K MEANS algorithms.” *Procedia Computer Science*, *176*, 1420-1429.
[14] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.
[15] Guo, Gongde, et al. "KNN model-based approach in classification." OTM Confederated International Conferences″ On the Move to Meaningful Internet Systems″. Springer, Berlin, Heidelberg, 2003.