Metamorphic Testing Approach for Verification of Sensor Relationships in Smart Spaces

Fahim Ahmed Irfan
Department of Computer Science
Missouri State University
Springfield, MO, USA
fai94s@missouristate.edu

Md Asif Tanvir
Department of Computer Science
Missouri State University
Springfield, MO, USA
mt5864s@missouristate.edu

Md Abdur Rahman Fahad
Department of Computer Science
Missouri State University
Springfield, MO, USA
mf8494s@missouristate.edu

***Abstract*— With the advancement of wireless technology, smart spaces such as smart homes, smart offices, etc. are increasingly equipped with smart devices such as physical sensors and actuators. In smart spaces, user activities produce time-series sensor data, which can be analyzed to derive rules for the autonomous activation of actuators such as smart lights, and smart switches. To achieve this, unsupervised learning techniques are employed, as the continuous growth of time-series data over time makes manual annotation challenging. However, the verification of the outputs of these unsupervised learning techniques often becomes challenging and complicated due to the lack of "ground truth" labels. Additionally, most unsupervised learning approaches have complex and ambiguous internal decision making mechanisms, making it difficult for end users to validate the output. Thus, these issues raise the "oracle problem", making it hard to validate and verify the unsupervised methods. In this paper, we present a metamorphic testing approach to verify the novel sensor grouping approach in smart spaces that leverages Spectral Clustering algorithm and graph-based feature representation from time series data. Our proposed approach consists of 10 metamorphic relations considering verification and validation aspects, with test cases designed according to those relations. The proposed approach entails test cases based on these relations, enabling the sensor grouping approach to endure the change in input data and variations in clustering parameters. Experimental results demonstrate that our proposed approach identifies the flaws in existing sensor relationship inference and verifies the outcomes of the clustering algorithm without any ground truth label, verifying the effectiveness of the clustering approach in diverse smart spaces. To validate the outcomes of metamorphic relations, we rely on unsupervised learning metrics such as Calinski-Harabasz Score (CH), Silhouette Score (SI) and Davies-Bouldin index (DB). Furthermore, we validate the correctness of sensor groups by manually checking each of the clusters generated after completing each metamorphic relation.**

Keywords—Adjacency matrix, Clustering, Metamorphic relation, Oracle problem, Sensor group

I. INTRODUCTION AND BACKGROUND

Automated solutions, including autonomous cars, industrial robotic automation, drone delivery, and AI-powered traffic control, have gained popularity with the advancement of artificial intelligence and the widespread use of physical sensors. With this trend, integration of smart assistants and physical sensors has increased in the domain of smart spaces such as smart homes, and smart offices, and with the number of users expected to reach approximately 785.16 million by 2028, according to Statista [1]. Consequently, there is a growing interest in automating home appliances interacting with smart assistants by setting up operational rules. To that end, the authors in [2] proposed an enforcement approach, while in [3], the authors demonstrated an operational policy that automates smart home actions based on sensor activities interacting with human behavior, utilizing an unsupervised learning approach. However, any undetected flaws in these automatic rule generations can affect the daily activities of home appliances, causing dissatisfaction to users. Therefore, it is important to uncover any flaws with rigorous testing so that the rule generation performs seamlessly within a software system.

Most machine learning solutions suffer from oracle problems in the absence of "test oracle" as determining the absolutely correct output for a given input becomes challenging, especially in complex and ambiguous situations. To overcome these oracle problems, metamorphic testing has emerged as a solution to effectively evaluating and testing machine learning approaches [4]. Researchers have utilized metamorphic testing in cancer diagnostic [5], image classifiers [6], and context recognition [7], etc. In [6], the authors experimented with two image classifiers: Support Vector Machine (SVM) classifier and Deep Learning based Convolutional Neural Network utilizing metamorphic relations. Their findings indicated that 71% of implementation bugs were resolved through metamorphic testing. While metamorphic testing proved effective for image classifiers, it also demonstrated efficacy in a context recognition model trained on textual information, as shown in [7]. The authors conducted 10 metamorphic relations (MRs) to validate the context recognition approach, one of which successfully identified a bug in the prediction model.

In the domain of unsupervised learning approaches, the efficacy of clustering algorithms like k-means [8], agglomerative clustering [9] and DBSCAN [10] in anomaly detection have been verified using metamorphic testing. In [9], the authors applied 14 metamorphic relations (MRs) to evaluate the performance of clustering algorithms, comparing K-means and agglomerative clustering. In [11], the authors used the metamorphic testing approach to automate the hyperparameters and features in any unsupervised method. However, most of these studies primarily assess algorithm efficacy in simulated environments or on custom pre-defined datasets, where outcome failures do not directly impact real-world scenarios. As unsupervised learning approaches lack ground truth labels, validating these methods in real-world automatic decision-making scenarios becomes essential. Therefore, in this paper, we propose a metamorphic testing approach to verify sensor grouping in smart spaces named SeReIn-M [12], utilizing an unsupervised learning approach. SeReIn-M leverages time-series sensor data to identify frequent sensor events that correspond to human activities, representing the features as an adjacency matrix. This matrix is then used as input for Spectral Clustering to cluster strongly connected sensor events. We propose 10 metamorphic relations considering two aspects of software testing:

Verification and Validation. Verification means the software is developed correctly without any error, while validation means the system is correct and meets client/user needs. Since there are no ground truth labels to verify this group, and the grouping may vary in different smart spaces, we selected metamorphic testing to verify the efficacy of SeReIn-M over any traditional software testing approach. Furthermore, during our literature review, we did not find enough research that conducted metamorphic testing to verify the outcomes of Spectral Clustering, which utilizes graph topology.

II. METAMORPHIC RELATIONS

We present a set of 10 MRs for our clustering algorithm. Each relation is designed to address either the verification or validation aspect of testing our approach on a sample time-series data. Verification-focused MRs assess whether SeReIn-M adheres to essential implementation characteristics, ensuring the algorithms functions as expected. In contrast, validation-focused MRs evaluate whether the SeReIn-M aligns with general user expectations. If there is no violation, SeReIn-M passes the test, otherwise it fails.

Given a source input *I*, where data instances (different sensor events) are assigned to clusters C_i (for i = 1, 2, 3, …., n). We denote the output of SeReIn-M for original dataset as *S* which denotes a set of clusters and *S′* is the output for the dataset after applying the corresponding MRs.

A. Metamorphic Relations for Verification

1) MR1. Adding Random Noise: MR1 assesses the robustness and efficacy of sensor grouping approach, SeReIn-M when randomly noise such as random sensor events are injected at random time. The hypothesis is that the original cluster should not be disrupted after adding random noise to the dataset, which means $S = S'$.

2) MR2. Duplicate Data Entries: MR2 assesses the stability of SeReIn-M, when duplicating the same sensor events randomly, creating multiple data instances of the same type. The hypothesis posits that the sensor grouping approach will accurately identify the same sensor groupings while disregarding the impact of duplication, which means $S = S'$.

3) MR3. Changing Affinity in Spectral Clustering: MR3 examines the effectiveness of SeReIn-M with different affinity such as precomputed, rbf, precomputed nearest neighbors, nearest neighbors while keeping the same data representation. The hypothesis posits that newly formed clusters will remain the same or exhibit a decline in quality, which means $S' \leq S$.

4) MR4. Scaling Features: The objective of MR4 is to identify the impact of scaling the extracted features with a multiplier M. The hypothesis suggests that the multiplier M will have no impact on generating clusters, meaning $S = S'$.

5) MR5. Reversing Data Order: This MR evaluates the behavior of sensor grouping approach, SeReIn-M, when the input data is provided with descending order. The hypothesis suggests that the sensor grouping approach will group similar sensors, meaning $S = S'$.

6) MR6. Partial Data Reversal: This MR examines the impact of SeReIn-M in the identification of sensor groups by reversing partial data of the given time series dataset. The hypothesis indicates that the sensor grouping approach will produce similar cluster or exihibit a decline in quality, meaning $S' \leq S$.

7) MR7. Data Reflection Effect: MR7 focuses on the impact of data reflection on SeReIn-M. We define data reflection as altering the binary states of sensors. The hypothesis suggests that there will be no impact of data reflection when new cluter is generated which means $S = S'$.

B. Metamorphic Relations for Validation

1) MR8. Removing Sensor: MR8 examines the relation inferencing capability once entire sensor is removed from a related sensor group containing at least two sensors, SeReIn-M still able to be related existing sensors. The hypothesis posits that the original cluster should remain same after removing any sensor from dataset, which means $S = S'$.

2) MR9. Adding New Sensor: MR9 examines the relation inferencing capability when a new sensor is added to a related sensor group maintaining similar time of event occurrence and SeReIn-M relates the added sensor with existing sensors. The hypothesis suggests that adding new sensors close to a sensor group should be clustered in same sensor group, producing same clusters which means $S = S'$.

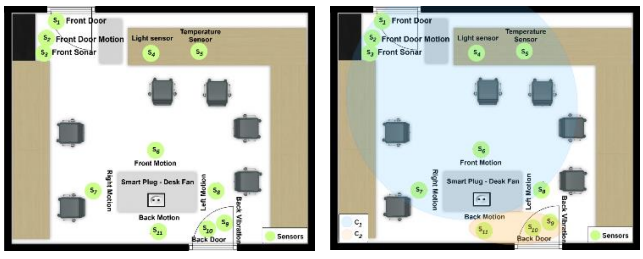
3) MR10. Small Data Volume Stability: MR10 evaluates the consistency of SeReIn-M in identifying relations among the sensor events within a small data proportion. The hypothesis implies that SeReIn-M will be able to identify the same relation with a subset of the dataset, meaning $S = S'$.

III. EXPERIMENT AND RESULTS

In this section, we will discuss the experimental setup along with the dataset. Finally, this section will be concluded by introducing the evaluation metrics and results generated by our MRs.

A. Experimental Setup

The To conduct our metamorphic relations, we relied on a smart office setup as shown in Figure 1a which was accessed by 5- 6 individuals, mostly on weekdays, and was equipped with motion, door, sonar, and vibration sensors to monitor activity, including room entry and exit. The dataset of this setup logged sensor events, defined as state changes at specific timestamps (in seconds). Therefore, the dataset includes the timestamp, sensor name, and sensor state, with each sensor assigned a unique identifier. Finally, utilizing this dataset, SeReIn-M that contains N-FNE and N-TD as feature extraction technique, produces sensor groups as shown in Figure 1b.



(a) Sensor Deployment (a) Actual Sensor Grouping

Fig. 1. Experimental Setup – Smart Office Layout

B. Evaluation Metrics

To measure the correctness of our proposed metamorphic relations, we employed CH score, SI score, and DB index—three commonly used metrics for assessing unsupervised clustering approaches. A higher CH score indicates stronger clusters, while the SI score ranges from -1 to +1, with positive values signifying well-defined clusters. The DB index, on the other hand, ranges from 0 to 1, where lower values indicate stronger clustering. We compare these metrics with the clustering scores received from before applying the metamorphic relations and after applying the relations. Table I shows the scores of base dataset before applying any relations.

TABLE I. CLUSTERING SCORES FOR BASE DATASET

Base Dataset					
N-FNE			***N-TD***		
CH	SI	DB	CH	SI	DB
21.02	0.70	0.32	16.37	0.64	0.45

C. Results

1) MR1. Adding Random Noise: We randomly added 25,000 noise data in the originial dataset. As for noise, we injected random values of sensor name, sensor state, event occurance to the dataset. Although the clustering scores decrease as an effect of adding noise, results presented in Table II shows that the sensor clusters identified after adding random noise to dataset are identical to the clusters generated by base dataset. Thus, we accept the hypothesis and SeReIn-M passes the test.

TABLE II. TESTING RESULTS FOR ADDING RANDOM NOISE

Adding Random Noise	
Base Dataset	Cluster 1: S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8
	Cluster 2: S_9, S_{10}, S_{11}
Adding Random Noise	Cluster 1: S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8
	Cluster 2: S_9, S_{10}, S_{11}
MR	Accept
Pass?	Yes

2) MR2. Duplicate Data Addition: We conducted experiments picking multiple random sensor data intance and adding 50,000 same data instance to the actual datase. Table III shows that after adding duplicate data instance, both N-FNE and N-TD have similar scores for all three metrics. Thus, we accept the hypothesis, confirming that SeReIn-M alings with expected outcome of MR2.

3) MR3. Changing Affinity in Spectral Clustering: We conducted the spectral clustering with different affinity such as precomputed, rbf, precomputed nearest neighbors, nearest neighbors. Table IV shows the clustering scores of SeReIn-M when we selected different affinity. As the score represents, in each affinity, the clustering scores are lower than the base clustering socre. Thus, we accept the hypothesis and SeReIn-M pass the test.

TABLE III. TESTING RESULTS FOR DUPLICATE DATA ADDITION

Result	**Duplicate Data Addition**					
	N-FNE			***N-TD***		
	CH	SI	DB	CH	SI	DB
	21.01	0.70	0.32	16.14	0.63	0.32
MR	Accept			Accept		
Pass?	Yes					

TABLE IV. TESTING RESULTS FOR CLUSTERING AFFINITY

Result	**Clustering Affinity**					
	N-FNE			***N-TD***		
	CH	SI	DB	CH	SI	DB
Rbf	2.71	0.09	1.61	2.11	0.05	7.71
Precomputed	21.01	0.70	0.32	16.14	0.63	0.32
Prcomputed nearest neighbors	0.57	-0.09	4.98	1.34	-0.13	1.58
Nearest neighbors	1.51	-0.07	1.79	3.17	0.16	1.47
MR	Accept			Accept		
Pass?	Yes					

4) MR4. Scaling Features: To test this MR, we multiplied the extracted features representing by adjacency matrix with constant M. Here, we multiplied the matrix with 50, 100 and 200. Table V shows that the clustering scores remained same even if the features were multiplied by constant M. Table IV shows that the clustering scores remained same even if the features were multiplied by constant M. Since each time the scores remained same, we showed one instance of the scores.

TABLE V. TESTING RESULTS FOR SCALING FEATURES

Result	**Scaling Features**					
	N-FNE			***N-TD***		
	CH	SI	DB	CH	SI	DB
	21.01	0.70	0.32	16.14	0.63	0.32
MR	Accept			Accept		
Pass?	Yes					

5) MR5. Reversing Data Order: Table VI shows the clustering scores when reversed the order of the dataset. Both N-FNE and N-TD remain same, proving the hypothesis MR5. Thus SeReIn-M passes this test.

TABLE VI. TESTING RESULTS FOR REVERSING DATA ORDER

Result	**Reversing Data Order**					
	N-FNE			***N-TD***		
	CH	SI	DB	CH	SI	DB
	21.01	0.70	0.32	16.14	0.63	0.32
MR	Accept			Accept		
Pass?	Yes					

6) MR6. Partial Data Reversal: We randomly reversed 30% of the base dataset. Table VII shows that for N-FNE the clustering scores improved while for N-TD declined. As per the hypothesis, both N-FNE and N-TD should perform lower than the base dataset. Therefore, we do not accept the hypothesis and SeReIn-M does not pass the test.

TABLE VII. TESTING RESULTS FOR PARTIAL DATA REVERSAL

Result	**Partial Data Reversal**					
	N-FNE			***N-TD***		
	CH	SI	DB	CH	SI	DB
	38.57	0.76	0.93	3.36	0.30	0.93
MR	Reject			Accept		
Pass?	No					

7) MR7. Data Reflection Effect: To test this MR, we alterned the binary sensor state. Specifically, if a sensor event was 1, we changed it to 0, and if it was 0, we changed it to 1, thereby creating a reflected version of the base dataset. Table VIII presents the clustering scores using the reflected dataset. The results indicate that the new clusters perform similarly to those in the base dataset. Therefore, we accept the hypothesis, confirming that SeReIn-M passes the test.

TABLE VIII. TESTING RESULTS FOR DATA REFLECTION

Result	**Data Reflection**					
	N-FNE			***N-TD***		
	CH	SI	DB	CH	SI	DB
	21.01	0.70	0.32	16.14	0.63	0.32
MR	Accept			Accept		
Pass?	Yes					

8) MR8. Removing Sensor: In this MR, we removed a sensor data instances from the base dataset that was grouped initially. In our experiment we removed Front Door Motion (S_2) from the dataset. As depicted in Figure 2 , after the removal of the sensor from the dataset, the remaining sensors still grouped in the same clusters. Thus, we accept the hypothesis and SeReIn-M pass the test.

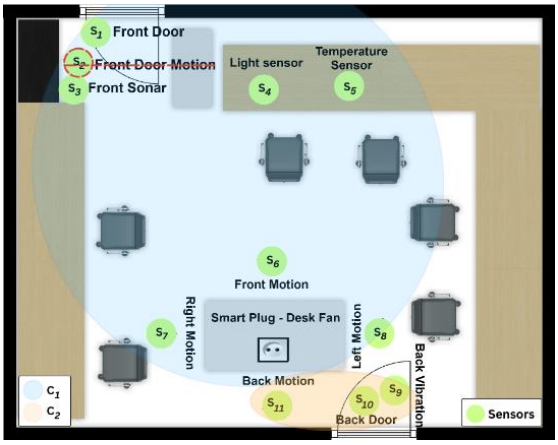


Fig. 2. Sensor Groups – Removing a sensor

9) MR9. Adding New Sensor: We experimented with SeReIn-M by adding new sensor data instances near the data instances of cluster C_2, as shown in Figure 3. Specifically, we introduced Back Sonar (S_{11}) into the initial cluster generated by the base dataset. As depicted in Figure 3, the newly added sensor grouped with the existing sensors of cluster C_2. Therefore, we accept the hypothesis, confirming that SeReIn-M passes the test.

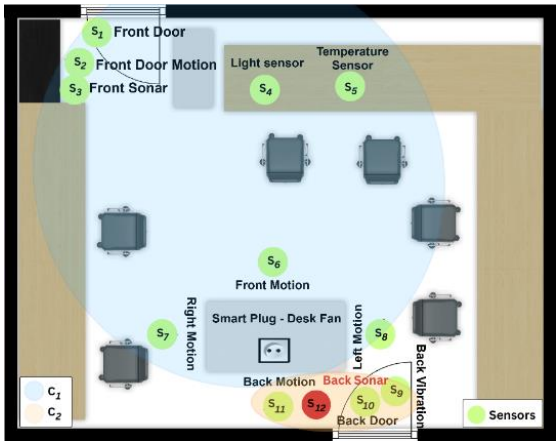


Fig. 3. Sensor Groups – Adding a sensor

10) MR10. Small Data Volume Stability: We experimented with 2 weeks of data (10% of the base dataset). As depicted in Figure 4, the clustering is identicial to the clustering obtained from base dataset. Thus, we accept the hypothesis and SeReIn-M passed the test.

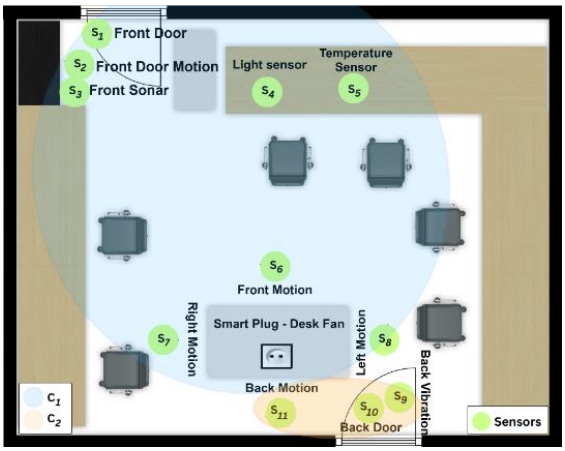


Fig. 4. Sensor Groups – Small Subset of Base Dataset

IV. CONCLUSION AND FUTURE WORK

In this project, we propose 10 metamorphic relations (MRs) to verify and validate the quality of sensor clustering in smart homes. Here, the first 7 MRs (MR1-MR7) are for verification purposes, and the last 3 MRs (MR8- MR10) are for validation purposes of the clustering method. We build these relations by modifying and altering some data parameters and some portions of the data. For verifying the outputs, we used clustering scores like CH score, SI score, and DB index. After our initial test, all of the relations passed, except the MR6. For that case, our hypothesis initially was proven wrong. Based on that MR, we debugged our code and fixed the code. After fixing, the MR passed,

thus proving the correctness of the clustering method. Also, for validation, we have generated the clusters using the MRs and manually plotted the sensor groups to confirm. Our initial inspection showcases that clustering method passes all the validation metamorphic tests also. For future work, this project can be extended to test on a more complex and bigger setup. Also, we have tried to cover most of the known metamorphic relations for unsupervised clustering. In the future, we plan to come up with some more robust and complex relations for the sensor grouping approach.

REFERENCES

[1] "Number of users of smart homes worldwide from 2019 to 2028," in Statista, 2025. [Online]. Available: https://www.statista.com/forecasts/887613/number-of-smart- homes-in-the-smart-home-market-in-the-worldJ.

[2] J. Hall and R. Iqbal, "Compes: A command messaging service for iot policy enforcement in a heterogeneous network," in Proceedings of the Second International Conference on Internet-of-Things Design and Implementation, 2017, pp. 37–43.

[3] S. P. Challa, R. Iqbal, and S. Liu, "An unsupervised learning approach for smart home operational policy generation," in Consumer Communi- cations Networking Conference (CCNC), 2023, pp. 1–6.

[4] F. U. Rehman and M. Srinivasan, "Metamorphic testing for machine learning: Applicability, challenges, and research opportunities," in Arti- ficial Intelligence Testing (AITest), 2023, pp. 34–39.

[5] S. H. Santos, B. N. C. Da Silveira, S. A. Andrade, M. Delamaro, and S. R. Souza, "An experimental study on applying metamorphic testing in machine learning applications," in Systematic and Automated Software Testing, 2020, pp. 98–106.

[6] A. Dwarakanath, M. Ahuja, S. Sikand, R. M. Rao, R. P. J. C. Bose, N. Dubash, and S. Podder, "Identifying implementation bugs in machine learning based image classifiers using metamorphic testing," in Software Testing and Analysis, 2018, p. 118–128.

[7] F. H. Athina, J. Sultana, D. D. Spandan, and R. Iqbal, "Metamorphic testing for investigation of context recognition from smart home voice commands," in 2024 7th Conference on Cloud and Internet of Things (CIoT), 2024, pp. 1–4.

[8] S. Yang, D. Towey, and Z. Q. Zhou, "Metamorphic exploration of an unsupervised clustering program," in Metamorphic Testing (MET), 2019, pp. 48–54.

[9] F. Ur Rehman and C. Izurieta, "Mt4uml: Metamorphic testing for unsupervised machine learning," in Swiss Conference on Data Science (SDS), 2022, pp. 26–32.

[10] F. U. Rehman and C. Izurieta, "An approach for verifying and validating clustering based anomaly detection systems using metamorphic testing," in Artificial Intelligence Testing (AITest), 2022, pp. 12–18.

[11] L. K. Shar, A. Goknil, E. J. Husom, S. Sen, Y. N. Tun, and K. Kim, "Autoconf: Automated configuration of unsupervised learning systems using metamorphic testing and bayesian optimization," in Automated Software Engineering (ASE), 2023, pp. 1326–1338.

[12] F. A. Irfan, R. Iqbal, and A. Siddiqua, "Serein-m: Sensor relationship inference in multi-resident smart homes," in Consumer Communications Networking Conference (CCNC), 2024, pp. 1–6.