해양 통신 시스템에서의 결측 데이터 문제와 진보된 데이터 대체 방법

신하 쉬르티카 [1], 염선호 [2], 박수현 [3*]
[1]국민대학교 금융정보보안학과
[2]국민대학교 금융정보보안학과
[3*]국민대학교 컴퓨터공학과

shrusinha09@kookmin.ac.kr, junsan86@kookmin.ac.kr, shpark21@kookmin.ac.kr

Challenges of Missing Data in Maritime Communication Systems and Advanced Data Imputation Methods

Shrutika Sinha[1], Sun-Ho Yum[2], Soo-Hyun Park[3*]
[1]Dept. of Financial Information Security, Kookmin University
[2]Dept. of Financial Information Security, Kookmin University
[3*]School of Computer Science, Kookmin University

요 약

A robust and efficient communication network is crucial for ensuring the smooth operation, efficiency, and utmost safety of various maritime activities. Throughout the vast expanse of history, the highly significant maritime industry has heavily relied upon an extensive range of diverse communication methods. Accurate network performance is critical in maritime environments, where data loss due to signal interruptions, equipment failures, and other domain-specific factors frequently occur. This paper evaluates traditional and advanced data imputation techniques to assess their impact on the predictive accuracy of machine learning models used for network switching decisions in maritime settings. Results show that advanced deep learning techniques, like Autoencoder-based imputation can improve performance over traditional methods.

1. 서론

Data is a vital component in every field, and its value is diminished when it is corrupt with incomplete information [1]. The maritime networks, in particular, are prone to gathering data that is often filled with gaps and inconsistencies [2]. Nevertheless, this data holds immense significance as it encompasses the intricate movements of contemporary vessels and various marine operations [3]. As a result, there is a growing imperative for the development of effective imputation techniques tailored specifically to the unique challenges presented by maritime network data [4]. It becomes evident that the application of diverse imputation methods across different networks becomes important in addressing these gaps. Hence, the incorporation of data imputation techniques is required in maritime networks [5]. Figure 1 is the representation data sharing in maritime operations, where various services like navigation, weather forecasting, search and rescue (SAR), warning, reporting and registration, and port info are provided.

Maritime networks are the most challenging when compared with land and satellite networks. They suffer from three unique environmental factors. First, the dynamic feature of the sea results in lower link longevity compared with that in land and satellite [6]. As a result, the observed traffic dynamics in the maritime environment may suffer from a larger amount of packet losses caused by link disconnections than those in the other environments [7]. Second, the characteristic temporal variation of signal strength in maritime radio channels contributes to the "signal instability" of maritime networks, which requires new types of temporal network modelling for these environments [8]. The third important factor that cannot be ignored is precipitation at sea may lead to higher path loss and serious fading when signals are transmitted through the atmosphere [9].

Maritime network data are faced with unique challenges [10]. They are often readily available, but they are also noisy and incomplete. These data are highly valued in many economic models measuring the importance of harbors as nodes of global production networks [11]. The significance of missing data on predictive and decision-making tools and

the comprehension of oceans across different domains and scales is widely recognized [12]. However, there is a necessity to prioritize actions in measuring these losses and creating efficient methods.

Therefore, to ensure data integrity in maritime operations, data imputation, which is a renowned data preprocessing mechanism, are introduced in this paper. This paper discusses some methods such mean, median, K-nearest neighbours (KNN) and Autoencoder imputation that can be applied to real-time datasets in maritime operations.

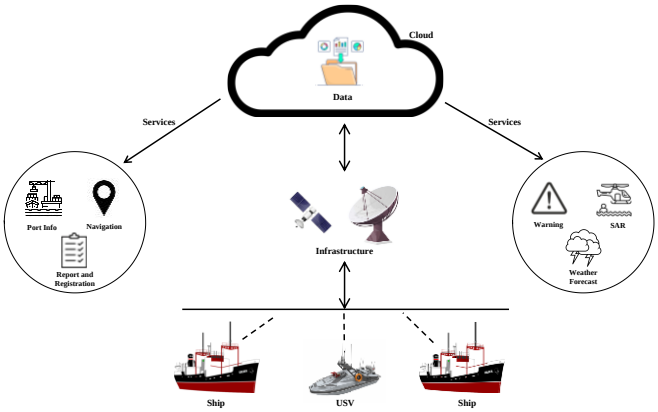

Figure 1. Data sharing in maritime operations.

The rest of the paper is organized as follows; Section II provides the methodology. Section III presents the results of the experiment. Section IV concludes the paper by summarizing the findings and outlining directions for future research.

2. Methodology

This section describes a study of missing data in a maritime communications system and details of the applicable data imputation methods. The example system consists of ships, USVs (Unmanned Surface Vehicles), and AUVs (Autonomous Underwater Vehicle) performing SAR operations in three dimensions at sea and underwater in the event of a life-threatening incident at sea. As a method of information exchange between equipment, a mobile communication network protocol that is highly reliable but requires network fees and an unlicensed broadband communication protocol that has high performance by design but decreases in reliability in adverse maritime conditions are switched according to the mission situation. The system continuously monitors the RX Rate (Reception Rate), TX Rate (Transmission Rate), RSSI Wi-Fi (Received Signal Strength Indicator for Wi-Fi), RX_CCQ (Reception Client Connection Quality), TX_CCQ (Transmission Client Connection Quality), RSSI LTE (Received Signal Strength Indicator for LTE), RSRP (Reference Signal Received Power), RSRQ (Reference Signal Received Quality), and SINR (Signal-to-Interference-plus-Noise Ratio) value of the primary communication protocol specified by the designer and has logic to switch to the protocol specified as the secondary communication if the value degrades to the point where packet exchange is not possible, and the results of the operation are recorded as a dataset. The expected missing data are assumed to be caused by the inability of the RSSI value, which serves as the switching trigger, to reflect the difficulties in predicting and measuring maritime communication systems described in the introduction.

To address the problem of missing data, several imputation techniques were applied to fill in the missing data and ensure the continuity of the dataset. The selected imputation methods include traditional techniques such as Mean, Median, and K-Nearest Neighbours (KNN) and an advanced deep learning approach, the Autoencoder.

A. Mean Imputation: Mean imputation is one of the simplest and commonly used techniques for imputing missing values. In this approach, the missing values of a feature are replaced by the mean value of that feature. It can be calculated separately for each feature and imputed with the mean value of that feature. In real-life applications, data can be missing because of various reasons. In practice, missing data can occur for a variety of reasons, such as technical failures leading to missing information during measurements, individuals intentionally not providing certain information, or values falling outside expected measurement ranges. Despite its effectiveness, the success of mean imputation depends on certain assumptions, such as the rate of missing values.

B. Median Imputation: Instead, median imputation uses the observed data to fill in missing values with a value that exists at least once within your variables: the median. This is especially helpful when the data being very skewed or contain outliers as the median tends to be far less affected by those extremes compared with mean. But again, like mean imputation, it does not consider relations between features and hence can be incorrect to a certain mechanism of missingness.

C. KNN Imputation: The K-nearest neighbours (KNN) imputation method is a widely recognized technique for filling in missing data based on finding similar observations. When an observation has missing values, the K nearest neighbours, or observations that most closely resemble it, are identified. The values from these similar observations are then used to fill in the missing values for the original observation. The process is repeated for all missing values. To determine the similarity between two observations, a distance metric is used, with the Euclidean being commonly utilized. This distance metric helps to identify the nearest neighbours. The KNN method is typically used after non-core variables have been removed, as it is specifically designed for filling in missing values for core variables.

D. Autoencoder: Autoencoder is a type of neural network that is used in unsupervised learning to learn concise, meaningful representations of the input data. The structure of an autoencoder resembles a symmetrical neural network. It consists of an input layer, an encoder, a decoder, and an

output layer. The input layer, also known as the visible layer, receives scaled data with features. The encoder compresses the data via two hidden layers: one consisting of 64 neurons with ReLU activation followed by another layer of only 32 neurons used as a memory bottleneck that summarises the main features needed for reconstruction. The decoder mirrors the encoder, restoring the compressed data through a hidden layer of 64 neurons which uses ReLU as an activation function and using linear again to match input dimensions. The output layer is responsible for rebuilding the input; therefore, we train our autoencoder to minimizes Mean Squared Error (MSE) loss between the original version of x and its reconstructed y that came from this one. This architecture has led the autoencoder to learn and infer missing values in maritime datasets well by taking non-linearity, and interrelations between features over time into consideration which is a rather powerful feature for preserving data integrity.

3. Results

This section presents the performance evaluation of different imputation methods—Mean, Median, KNN, and Autoencoder—on the maritime dataset, with a focus on accurately predicting Wi-Fi and LTE connections

The accuracy results indicate that Autoencoder imputation methods significantly outperform the traditional statistical methods. Autoencoder achieved the highest accuracy of 98 %, demonstrating their superior capability in handling missing data in this context. Mean, Median and KNN imputations yielded an accuracy of 95% and 97%, showing reasonable performance but failing to capture the complexities in the data as effectively as the autoencoder. Figure 2 and Figure 3 represent the performance of all the metrics for Wi-Fi and LTE respectively. 3D contour plots representing the performance of various imputation methods (Mean, Median, KNN and Autoencoder) over Precision Recall and F1-Score for Wi-Fi and LTE Data. The x-axis displays the imputation methods, y-axis represents metrics and z axis shows scores from 0.80 to 1.00. Peaks in the plots indicate higher performance, particularly for KNN and Autoencoder methods, which consistently achieve the best scores across all metrics. In contrast, the Mean and Median methods show lower performance with more variability, as represented by the lower contour regions.

TABLE I. ACCURACY FOR EACH IMPUTATION METHOD

Method	Accuracy
Mean Imputation	0.95
Median Imputation	0.95
KNN Imputation	0.97
Autoencoder Imputation	0.98

The results summarized in Table 1, which details the accuracy, for each imputation method, highlighting their effectiveness across Wi-Fi and LTE. The accuracy results indicate that Autoencoder imputation methods significantly outperform the traditional statistical methods. Autoencoder achieved the highest accuracy of 98 %, demonstrating their superior capability in handling missing data in this context. Mean, Median and KNN imputations yielded an accuracy of 95% and 97%, showing reasonable performance but failing to capture the complexities in the data as effectively as the autoencoder.

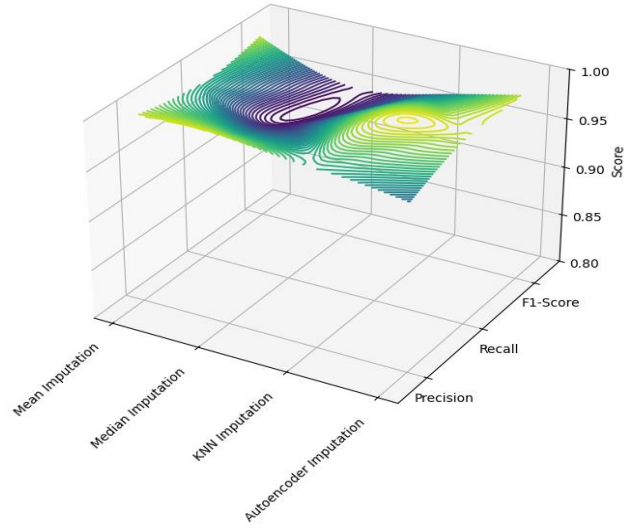

Figure 2. Wi-Fi Performance Metrics by Imputation Method.

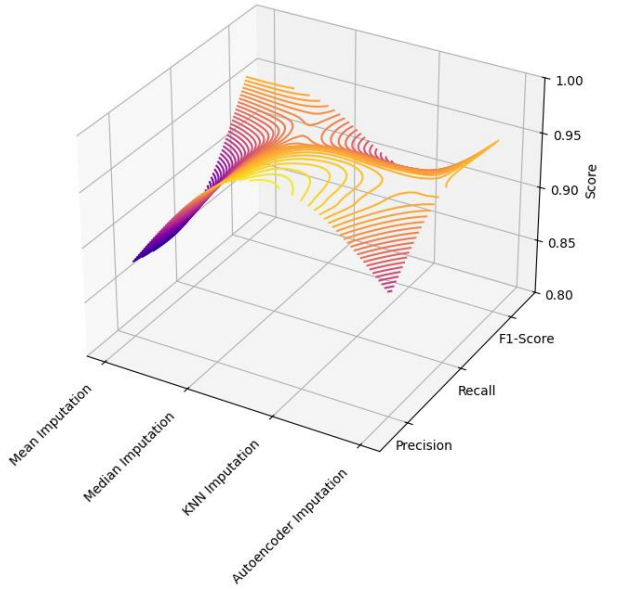

Figure 3. LTE Performance Metrics by Imputation Method.

4. Conclusion

This paper highlights the important role of advanced imputation techniques in improving classification performance in real-world datasets with missing values. Autoencoders, has balance of accuracy and adaptability, making them an excellent tool for data-intensive maritime research. Future studies could explore the integration of other deep learning architectures, such as Variational Autoencoders or Generative Adversarial Networks, to further enhance imputation performance. Additionally, applying these

imputation techniques to larger and more diverse maritime datasets could provide deeper insights into their scalability and generalizability across different maritime applications.

참고문헌

[1] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. C. Guttula, S. Mujumdar, S. Afzal, R. S. Mittal, and V. Munigala, “Overview and importance of data quality for machine learning tasks,” Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2020. doi: 10.1145/3394486.3406477.

[2] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, “Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems,” IEEE Internet Things J., vol. 8, pp. 5374–5385, 2021. doi: 10.1109/JIOT.2020.3028743.

[3] Z. Peng, J. Wang, D. Wang, and Q. Han, “An overview of recent advances in coordinated control of multiple autonomous surface vehicles,” IEEE Trans. Ind. Informat., vol. 17, pp. 732–745, 2021. doi: 10.1109/TII.2020.3004343.

[4] G. Wang, M. Ma, L. Jiang, F. Chen, and L. Xu, “Multiple imputation of maritime search and rescue data at multiple missing patterns,” PLoS ONE, vol. 16, 2021. doi: 10.1371/journal.pone.0252129.

[5] Y.-y. Chen, Y. Lv, and F.-y. Wang, “Traffic flow imputation using parallel data and generative adversarial networks,” IEEE Trans. Intell. Transp. Syst., vol. 21, pp. 1624–1630, 2020. doi: 10.1109/TITS.2019.2910295.

[6] K. Yau, A. R. Syed, W. Hashim, J. Qadir, C. Wu, and N. Hassan, “Maritime networking: Bringing internet to the sea,” IEEE Access, vol. 7, pp. 48236–48255, 2019. doi: 10.1109/ACCESS.2019.2909921.

[7] S. Göksu and Ö. Arslan, “Quantitative analysis of dynamic risk factors for shipping operations,” J. ETA Marit. Sci., vol. 8, no. 2, 2020. doi: 10.5505/jems.2020.63308.

[8] P. Bithas, E. T. Michailidis, N. Nomikos, D. Vouyioukas, and A. Kanatas, “A survey on machine-learning techniques for UAV-based communications,” *Sensors (Basel, Switzerland)*, vol. 19, 2019. doi: 10.3390/s19235170.

[9] S. Sinha, G. P. Reddy, and S.-H. Park, "Channel selection using machine learning," 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 2024, pp. 164-168. doi: 10.1109/ICAIIC60209.2024.10463321.

[10] G. P. Reddy, S. Sinha, and S.-H. Park, “Analysis of research trends in maritime communication,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 6, 2024. doi: 10.14569/ijacsa.2024.0150670.

[11] S. Sinha, S. Y. Kwon, and S. H. Park, “A survey of deep/machine learning in maritime communications,” 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), Paris, France, 2023, pp. 856-860. doi: 10.1109/ICUFN57995.2023.10199678.

[12] T. Bolton and L. Zanna, “Applications of deep learning to ocean data inference and subgrid parameterization,” *J. Adv. Model. Earth Syst.*, vol. 11, pp. 376–399, 2019. doi: 10.1029/2018MS001472.

ACKNOWLEDGEMENT

이 논문은 2024 년도 해양경찰청 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임(RS-2021-KS211488, 해양사고 신속대응 군집수색 자율수중로봇시스템 개발)