Feature Extraction and Clustering of High Dimensional Electromagnetic Interference Signals Based on Multidimensional Scaling and SOM Network

Hongyi Li¹, Tianshi Xu¹, Haoran Lian¹, Wei Chen² and Di Zhao¹,*

¹ LMIB, School of Mathematical Sciences, Beihang University, Beijing, China
*Corresponding author’s e-mail: zdhyl2010@163.com

Abstract. With the wide application of communication technology, the threat of interference is becoming more and more serious. These disturbances usually appear as higher-dimensional vectors, and the nuances between the different vectors are indistinguishable to humans. A great deal of research has been done on how to deal with this interference in an automated way. But most of the computation time in these studies was spent on the training and computation of the convolution layer. As the electromagnetic equipment tends to be a large inheritance system, the electromagnetic signal also presents higher dimensional characteristics. Then the convolutional neural network represented must use larger and more convolutional layers, so the time cost is very high. Therefore, feature extraction of the original signal is carried out before clustering. After adjusting the selection of features, the algorithm performs well on our dataset.

1. Introduction
With the development of technology, the electromagnetic equipment gradually develops to a large-scale integrated system. Therefore, the EMI signals exhibit high-dimensional, strong nonlinear, and non-stationary characteristics[1]. This paper focuses on clustering applications in EMC analysis. Some measures for separating the signals have been proposed[2,3] they do work, but these measures often require huge computational resources and a lot of time.

In the literature [2], most of the computational time is spent on the calculation of convolutional layers. When faced with higher dimensional and larger electromagnetic data, convolutional neural networks must have more parameters which cost a great deal of time. But in fact, there may be a strong correlation between the basis functions composed of the original high-dimensional features. Therefore, it is necessary to optimize first the original base function system, and reduce the correlation between the base functions as much as possible while ensuring the completeness of the base function library to minimize the time overhead of clustering.

While on clustering algorithms, there are many general algorithms, including K-means, FCN, etc[4]. But they have the same disadvantages: they do not exploit the feedback efficiently. To solve all these problems, we proposes a novel method consisting of multidimensional scaling and SOM networks, first performing dimension reduction while maintaining the signal features, and then clustering using SOM networks. Experiments show that we reduce the signal length by 90%, but still maintain a high accuracy, and even improve the various indicators.
2. Materials and Methods

2.1. Signal dimension reduction technology

2.1.1 Technical Overview
To eliminate the dispensable features to reduce the computational overhead while solving the dimensional disaster problem, a dimensionality reduction of the high-dimensional signals is required. There are generally two ways to solve these problems:

1) feature selection. Feature selection performs the feature extraction task only on the associated subspace. It usually uses greedy policies to search for different characteristic subspaces, and then uses some criteria to evaluate these subspaces. However, this method requires manual extraction of characteristics and considerable professional knowledge, time-consuming and force consuming.

2) feature transformation, this approach includes many traditional methods, such as principal component analysis (PCA), Laplace map (LE), isometric mapping (ISOMAP) and multidimensional scaling (MDS). PCA maintains the maximum variance between data and focuses on preserving the maximum separability of the data; LE main idea is to preserve the structure between data as constant as possible in low dimensional space[5]; ISOMAP is actually a variant of MDS, but it uses a geodesic distance, not an Euclidean distance[6]; MDS keeps the "distance" between the data unchanged, pays more attention to the features inside the high-dimensional data, and retains the "similarity" information in the high-dimensional space, which is more conducive to clustering high-dimensional signal[7].

2.1.2 multidimensional scaling (MDS)
Multidimensional scaling, a linear dimension reduction approach, different from both principal and linear dimension reduction, the goal of multidimensional scaling is not to preserve the maximum separability of the data, but to have more focus on features internal with high-dimensional data. Multidimensional scaling algorithms focus on preserving "similarity" information in high-dimensional spaces, which is often defined by Euclidean distance during general problem resolution.

2.2 Self Organizing Maps (SOM)
SOM neural network assumes some topology or order in the input object that can implement a dimensional reduction mapping from the input space (n dimensions) to the output plane (2 dimensions) with topological feature preserving properties and strong theoretical connections with actual brain processing[8]. Self-organized mapping neural networks allow clustering of data from unsupervised learning. Its idea is simple and essentially a neural network with only input and hidden layers. A node in the hidden layer represents a class that needs to be gathered. Training in a "competitive learning" approach, each input sample finds a node it most matches in the hidden layer, called its activation node, immediately followed by updating the parameters of the activation node with stochastic gradient descent. Meanwhile, points and the activation node closely update parameters according to their distance to the activation node. In fact, previous work by the researchers shows that SOM networks have great potential for clustering time series data and in some cases can provide more accurate results than other clustering algorithms. Its network structure is shown in the following figure:

![Figure 1 SOM Network Structure](image)
2.3 Proposed model
In the previous section, we introduce signal dimension reduction techniques and SOM theory. This paper presents a high-dimensional electromagnetic interference signal clustering method based on MDS dimension reduction techniques and SOM neural networks, where the SOM neural network adopts 24 * 24 gauge size and the number of clusters is 34, ultimately providing an accurate mathematical simulation model for the design, analysis, prediction and evaluation techniques of EMC.

3. Simulation Experiment
In this section, we conduct experiments using the above methods on the simulated signal dataset. First, we will introduce the signal dataset used in the experiment, Section 3.1. We then discuss the metrics used in the experiment in section 3, Section 3.2. Final experimental results are shown in Section 3.3.

3.1 Analog signal generation
The data used in the experiment was generated through several noise signal generator programs written by Exstrom Labs, LLC. The dataset contains five signals, pink (P), exponential (ED), Laplace (LD), uniform (UD) and Brownian motion (BM), and we randomly extract a sample from each noise, as shown in Figure 2-3. The P noise is characterized by the power spectral density inversely proportional to the frequency. The ED,LD,UD noise is generated by random sampling according to a certain distribution. BM noise is generated by sampling and a complete Gaussian stochastic process. These noises are essentially very close to electromagnetic signals and are therefore used as analog electromagnetic signals. In the dataset, 1000 signals for each noise, totaling 5000 signals constitute the training set and test set of this experiment.

3.2 Evaluation indicators
We measure the results of our clustering using the following metrics:

1) Accuracy(ACC). Accuracy demonstrates how many samples are properly assembled into a group.

2) Adjusted Rand Index (ARI)[9]. ARI is an indicator of measuring the similarity between clustering and real results that ignores permutations and reflects the degree of overlap between clustering and real results. The value of ARI is between -1 and 1 the closer 1 the better the performance.

3) Normal mutual information (NMI)[10]. Used to measure the similarity of the two clustering results, and it encourages the number of clusters to be as small as possible. NMI values between 0-1, the closer 1 the better the performance.

4) V-Measure[11]. This index covers two indicators: homogeneity and completeness. High
homogeneity indicates that each cluster contains only members of one class, while high integrity indicates that all members of a given class are assigned to the same cluster. The V-Measure is between 0 and 1. The greater this value, the better the performance.

3.3 Experimental results and analysis

3.3.1 experiment 1: No extracting features
In experiment 1, we do not make any feature extraction and dimension reduction, only clustering the original 500-dimensional high-dimensional signals as input. After 200 coarse training iterations, 30 fine training iterations are shown in Table 1:

| Experimental results |
|----------------------|
| ACC  | ARI  | NMI  | V-Measure |
|-----|------|------|-----------|
| Experimental 1       | 0.7984 | 0.8401 | 0.8237 | 0.9954 |

3.3.2 experiment 2: Using feature extraction
In Experiment 2, we use Principal Component Analysis, Laplace Map, Isometric Map and Multidimensional Scaling respectively to reduce the dimension of the original data from 500 to 50, which reduces by 90% and preserves the most critical features of the signal. The other conditions are equivalent to experiment 1. After 200 rough training iterations, the results of 30 fine training iterations are shown in Table 2:

| Experimental results |
|----------------------|
| ACC  | ARI  | NMI  | V-Measure |
|-----|------|------|-----------|
| Experimental 1       | 0.7984 | 0.8401 | 0.8237 | 0.9954 |
| PCA | 0.7358 | 0.7352 | 0.74035 | 0.9294 |
| LE  | 0.2136 | 0.1970 | 0.5455 | 0.8521 |
| ISOMAP | 0.3626 | 0.3188 | 0.5721 | 0.5890 |
| MDS | **0.8074** | **0.8615** | **0.8237** | **0.9959** |

It can be seen that all indicators of PCA feature extraction have decreased to a certain extent, indicating that PCA is not effective in extracting such signal features and some key features will be lost. With LE and ISOMAP, all indicators declined significantly, indicating that most of the key features are lost. Finally, using MDS method, all indicators did not decrease, but improved, and the V-Measure index reached 0.9959, indicating that the features extracted by the model reflect the overall characteristics of the signal very effectively. The results show that the feature extraction using multi-dimensional MDS can significantly reduce the dimension of the signal, eliminate the redundant basis function, and maintain the ability to describe the complete signal feature.

3.3.3 Experiment 3: Compared with other algorithms
In Experiment 3, we compare our algorithm with other known algorithms on the same dataset, with the results as shown in Table 3:
Table 3 Results of experiment 3

| Experimental results | ACC   | ARI   | NMI   | V-Measure |
|----------------------|-------|-------|-------|-----------|
| MDS+SOM              | 0.8074| 0.8615| 0.8237| 0.9959    |
| FastICA[12]+ SOM     | 0.774 | 0.809 | \    | 0.858     |
| TruncatedSVD[13]+ SOM| 0.787 | 0.811 | \    | 0.864     |

It can be seen that also using SOM clustering, using MDS is significantly better than the other two, and note that we also reduced the dimension by 90%, which further proves that MDS is more effective for feature extraction of such signals.

4. Conclusions

In this paper, we propose a model that combines feature extraction with clustering, among them, multidimensional scaling performs well in this problem. Compared to previous studies, the signal length after feature extraction processing is reduced by 90% and is able to well preserve the characteristics of the original signal itself. In addition to the improvement of the various indicators of the clustering results, the time of training the model is also greatly shortened. Moreover, the model has good scalability. With the gradual development of electromagnetic equipment to large-scale integration, the length and complexity of signals are also constantly improving, and the method based on feature extraction will inevitably be more widely used.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant no. 61771001).

References

[1] Albertson, R. T., Arthur, J., & Rashid, M. H. (2006, September). Overview of electromagnetic interference. In 2006 38th North American Power Symposium (pp. 263-266). IEEE.
[2] Hongyi Li, Wei Chen, Wei Lin, and Di Zhao. (2019) Online and Semi-Supervised Clustering of Electromagnetic Interference Signals with Convolutional Layers. Access IEEE.
[3] Hongyi Li, Di Zhao, Shaofeng Xu, Pidong Wang and Jiaxin Chen. (2016). High Dimensional Electromagnetic Interference Signal Clustering Based On SOM Neural Network. ELECTRONICS, VOL. 20, NO. 1.
[4] Shelhamer Evan, Long Jonathan, Darrell Trevor. Fully Convolutional Networks for Semantic Segmentation. (2017) IEEE transactions on pattern analysis and machine intelligence, 39:4-4.
[5] Belkin M, Niyogi P. (2014) Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. MIT Press.
[6] Tenenbaum J B, Silva V D, Langford J C. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science, 290(5500): 2319-2323.
[7] Young, Reynolds, Lance M, et al. (1981) Introduction to multidimensional scaling.: Academic Press.
[8] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.
[9] Hubert, L. and Arabie, P. (1985) Comparing partitions. Journal of Classification, 193–218.
[10] Estevéz P A, Tesmer M, Perez C A, et al. (2009) Normalized Mutual Information Feature Selection. IEEE Transactions on Neural Networks, 20(2): 189-201.
[11] Rosenberg, A., & Hirschberg, J. (2007, June). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 410-420).
[12] Wang, M. C., & Uhlenbeck, G. E. (1945). On the theory of the Brownian motion II. Reviews of modern physics, 17(2-3), 323.

[13] Oja, E., & Hyvarinen, A. (2000). Independent component analysis: algorithms and applications. Neural networks, 13(4-5), 411-430.