

Study of Validation Methods for Augmented Dital Evidence Data

Jong-Jin Jung*, Jong-Bin Park**

*(Information Media Research Center, Korea Electronics Technology Institute)

** (Information Media Research Center, Korea Electronics Technology Institute)

ABSTRACT

This paper introduces a study to verify whether the expanded data through various data augmentation methods are valid in terms of accuracy and bias. Data augmentation is a method of processing and generating other types of data with similar characteristics based on the characteristics of the obtained data, rather than directly collecting data when there is not enough data to increase analysis accuracy. However, unverified and augmented data may actually degrade the results of the analysis. Before using the amplified data for analysis, it is a very important verification factor whether it is accurately propagated in terms of similarity to the source data, and whether bias occurs because only a specific part is concentrated and propagated as a result of the propagation. Therefore, in this paper, a verification method is presented from these two perspectives.

Keywords - data augmentation, validation, biased distribution, embedding vector

Date of Submission: 01-07-2023

Date of acceptance: 11-07-2023

I. INTRODUCTION

With the development of various industries, new technologies and platforms, new artificial intelligence models are constantly appearing in the market, and how much high-quality data can be secured and analyzed well to provide accurate information and services is becoming a core competitiveness of business. By classifying the obtained raw data, metadata with different shapes (text, image, audio, video, etc.) Securing high-quality data through multiplication or fusion between data can increase the accuracy of analysis[1]. In addition, the time and effort required for data preparation will be dramatically reduced by self-replicating the data required for AI learning rather than relying solely on collection. For example, first, if video data is secured as seed data, co-acoustic, voice, image, text, etc. are respectively extracted from there, and only items having a common subject are selected. Fig. 1 briefly shows this process. It is a method to obtain extracted and selected information of different types through open data sharing portals (CoCo ImageNet, Kaggle, AI Dataset Hub) that provide similar data. However, not all physically increased data improves the accuracy of analysis[2]. That is, data similar to the source data used for propagation and accurate data should be propagated, and the results of propagation should not be biased in terms of the distribution of the secured data so as not to lead to analysis errors. Therefore, in

this paper, we examine the research using only data that can improve the results of analysis by verifying the augmented data in terms of validity and bias.

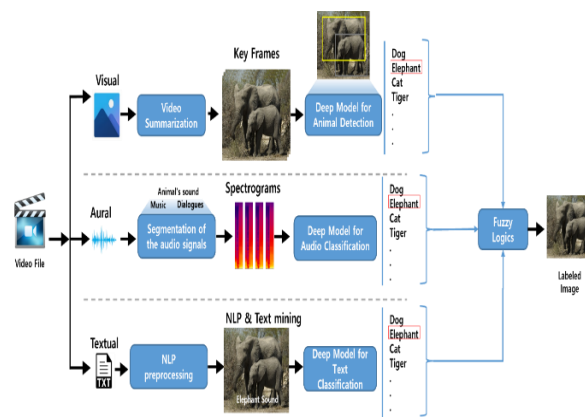


Figure 1. Example of data extraction and augmentation process

II. VALIDATION OF AUGMENTED DATA

A. Validation Overview

Data analysts should be able to obtain information about whether the augmented data meets the purpose of my analysis through validation. Through the validation results, it should be possible to check whether the bias is improved when expanding to the data storage for final analysis with the target of multiplication data that meets the

purpose of analysis. The goal of this study is to extend the amplified data to the final data storage by selecting data with outliers as a result of specific validation after checking whether the accuracy and bias are improved for the augmented data.

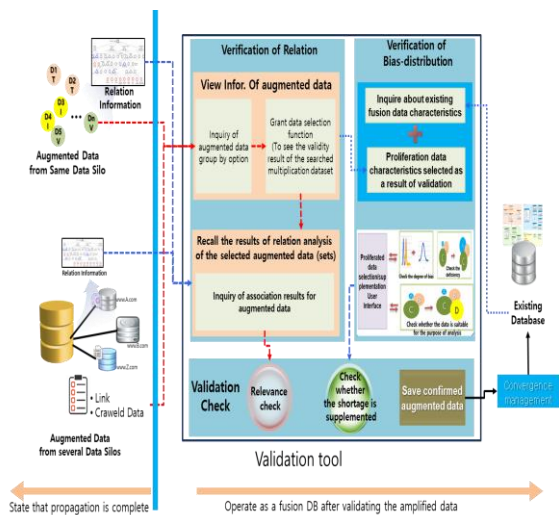


Figure 2. Validation and final data management schematic

According to Fig. 2, when the user selects the seed data requested for propagation, the corresponding propagated data list is checked. Identify the augmented evidence (association analysis results) of the augmented data presented in the list. After confirming the relation analysis result, that is, the reason for proliferation, the validation result derived according to the validation model is finally confirmed. Only high-quality, proliferated data with more than a specific validation value is selected and managed as a learning data extension DB. The validation tool provides validation information in terms of 1) data accuracy and 2) improvement of the bias of the expanded data for the augmented data by the data analyst[3]. In addition, the validation tool also provides a convenient inquiry/search function to expand the data required for the desired analysis by quickly and easily searching the augmented data. The main components and functions of the validation tool proposed in this paper are shown in Table 1 and Fig. 3[4].

Table 1: DETAILED STUDY BY VALIDATION ITEM

Items	Detailed study content
Verification of correlation between seed data and proliferation data	<ul style="list-style-type: none"> • Perform high-dimensional vector space embedding. • Calculation of relation in embedded vector space • Clustering and outlier removal • Dynamic association verification

Items	Detailed study content
	considering prior knowledge based on transfer learning
Bias analysis and data reprocessing	<ul style="list-style-type: none"> • Bias analysis based on low-level meta information • High-level bias analysis based on high-dimensional meta information • Core technology for securing new data sets through variable-controlled data reprocessing

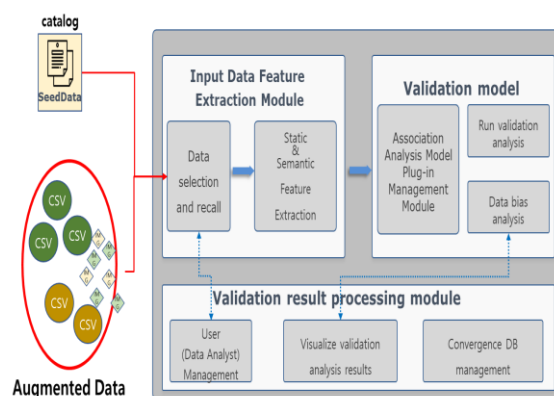


Figure 3. Task process flow diagram of validation tool

B. Validation

As mentioned in the previous explanation, in this paper, it is judged whether the amplified data can be used for analysis in terms of accuracy and bias. Fig. 4 shows the concept and specific validation methodology for text data and image data[5].

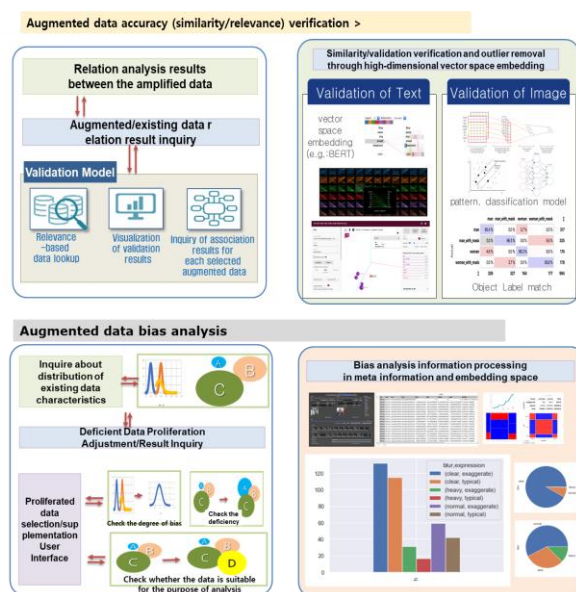


Figure 4. Validation details by verification item

C. Detailed validation process

The model that performs validation on input seed data and vertical and horizontally propagated data is defined as follows.

Table 2 : Validation Model Design

<ul style="list-style-type: none"> - Input Seed Data : S - Extract data set space : $\{A_1, A_2, A_3 \dots A_n\}, \{n: n \in N\}$ - Augmented data set : $A = \begin{Bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ A_{21} & A_{22} & \dots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nm} \end{Bmatrix}$ - A_{nm} is an augmented data element, belonging to the set $\{A_1, A_2, A_3 \dots A_n\}, \{n: n \in N\}$. It means that the m^{th} element $\cdot \cdot A_n$ is created. - For concise expression, it is shown that m-number multiplications are performed equally for each n elements, but the number of data generated for each element may vary. - The similarity calculation can be defined as the process of calculating the relevance of the extracted data A_i and A_{ij}. - Validation can be defined as a process of verifying whether the propagated data is valid. It is desirable to determine Acceptance and Rejection by checking whether the calculated similarity value is larger or smaller than the user-defined threshold T.

D. Metrics for calculating similarity for validation.

For similarity calculation, it is desirable to use a metric suitable for each characteristic for the same media type for images, audio, and text, but metrics such as LSA, LDA, and Word2Vec can be used for comparison between heterogeneous types[6]. For example, by extracting high-level meta information (Semantic Information) from each input data and measuring the similarity in the corresponding meta information space, it is possible to verify the validity of different types of data. MSE (Mean Squared Error): It is mainly used for comparison between two-dimensional images, but it is applicable to a one-dimensional audio signal, and in the case of a video, it can be regarded as a set of images, audio, and text, so it is a partially applicable metric. 0 means perfect match[7].

III. SIMULATION RESULTS

E. Process of simulation tasks

Accuracy verification is a procedure that verifies how similar the source data used for propagation and the propagated data are. In other words, it is to exclude proliferation data with poor relevance from the results of proliferation. The

validation procedure used in this study is shown in Fig 5.

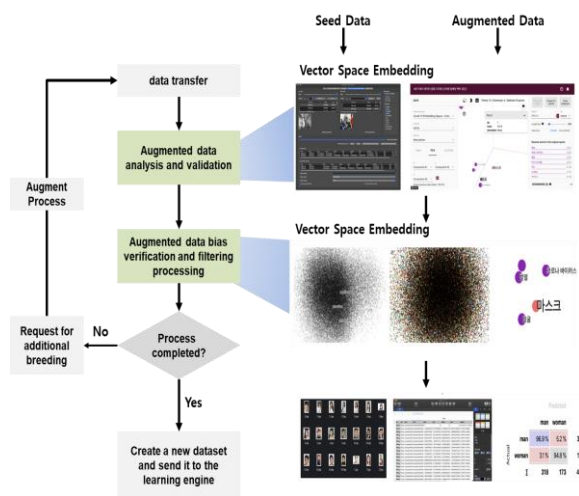


Figure 5. Validation task process and its simulation result

F. Bias validation and data reconstruction

By analyzing both the data increased as a result of the proliferation and the existing data, the distribution according to the data characteristics is examined. As a method of determining bias, basic bias analysis and verification are performed by analyzing low-level meta information such as data type, size, and length. Secure an extended dataset from which bias is removed.

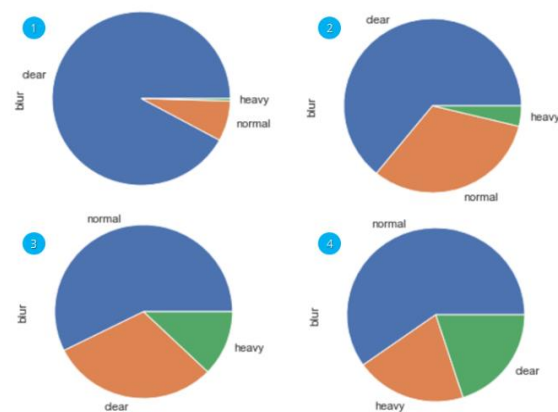
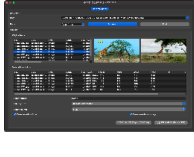


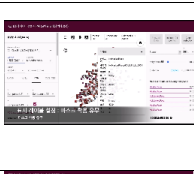


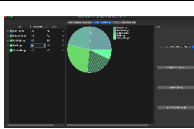
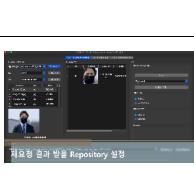


Figure 6. Example of improving bias check and distribution reconstruction

G. Visualized Scenes of Validation Tool

In this chapter, we present the scene results to verify the data validity through the verification tool presented in this paper. Table 3 show the main propagated data validation scenes.

Table 3. Detailed study by validation item

Validation Scenes	Detailed content
	<ul style="list-style-type: none"> • Accuracy of the amplified data target (comprehensive query of basic association analysis information)
	<ul style="list-style-type: none"> • Check the feature in which the propagated information is Vector Embedding
	<ul style="list-style-type: none"> • Check the embedding coefficient (Tensor) value where the extended features are located in the vector space
	<ul style="list-style-type: none"> • Provides a user interface that enables information accuracy inquiry and selection of data to be propagated through various options (embedding features obtained through analysis)
	<ul style="list-style-type: none"> • Provides a heuristic information transfer interface through dimension reduction
	<ul style="list-style-type: none"> • Provides an interface to check outlier data between clusters (check individual cluster center points and outliers from distance thresholds)
	<ul style="list-style-type: none"> • Development of an interface to check the feature distribution of the multiplied data
	<ul style="list-style-type: none"> • Through the above results, it is possible to indirectly review data bias • Provides an interface for specifying the propagation request and completed data to the propagation engine along with the desired data distribution information.

IV. CONCLUSION

In this paper, we introduced the concept of heterogeneous data propagation that can easily and quickly obtain data for AI learning. We looked at the key technologies to realize this. Currently, the method of quantitative expansion of learning data that is mainly used is mainly data growth targeting only the same type of data, so it does not meet the demand for complex data analysis.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2023-00225661, Development of causal reasoning and expression technology to enhance the proof of digital evidence)

REFERENCES

- [1]. Liwei Wang, Yin Li, Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings", Proceedings of the IEEE conference on computer vision and pattern recognition, pp.5005—5013, 2016
- [2]. Deerwester, Scott C., et al. "Indexing by latent semantic analysis." JASIS 41.6 pp.391-407, 1990
- [3]. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3, pp. 993-1022, 2003
- [4]. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, , "Rethinking the inception architecture for computer vision, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818--2826, 2016
- [5]. Yijun Li, Sifei Liu, Jimei Yang, Ming -Hsuan Yang, " Generative face completion", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3911-3919, 2017
- [6]. Embedding projector, <https://github.com/tensorflow/embedding-projector-standalone>.
- [7]. M. Jordan, T. Mitchell, "Machine learning: Trends, perspectives, and prospects", Science, vol. 349, no. 6245, pp. 255-260, 2015