A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process

Dongting Xu[a,b], Zhisheng Zhang[a], Jinfei Shi[a,b]
[a] *School of Mechanical Engineering*, *Southeast University, Nanjing, China*
[b] *School of Mechanical Engineering, Nanjing Institute of Technology, Nanjing, China*
xudongting@seu.edu.cn, oldbc@seu.edu.cn, shijf_xdt@163.com

2022 5th International Conference on Data Science and Information Technology (DSIT) | 978-1-6654-9868-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSIT55514.2022.9943883

***Abstract*—Data can be viewed as a product. The quality of data can be assessed as the quality of a product which can be evaluated by multiple dimensions. Traditional data quality only refers to accuracy in data collecting system. In the AI age, the accuracy is no longer the only dimension for measuring data quality. Data quality has been expanded into sophisticated comprehensive concept. For the same set of data, different users may have different standards for the data quality. The high quality of data is a prerequisite for effective modeling and obtaining value from data. A data quality assessment matrix composed by multiple dimensions can measure the data quality subjectively and objectively. For a particular application, the data quality assessment matrix should reflect the degree of satisfaction of the data. In this study, a data assessment matrix for imbalanced multivariate time series data from complex manufacturing process is designed to measure the data quality quantitatively. Multiple sensors are placed along the machines and the process for collecting signals about the manufacturing process can be viewed as a data production process. Deep learning methods can be applied to the data to monitor the machines and predict the paper break failures. A control chart is designed for the data production process to control the data quality and help to improve the quality of training dataset for deep learning models. The control chart is more directly than hierarchical clustering result. The result of assessment and the control chart could help the data users to understand the data deeper and help build better models.**

Keywords—data quality, data production process, subjective and objective assessment, imbalanced multivariate time series data, control chart, training data quality improvement

I. INTRODUCTION

Recent achievements in science and technology have enabled the growth and availability of data acquisition and collection. Knowledge discovery and data engineering are playing an important role in a wide range of areas. Data is in a powerful fashion as the development of Artificial Intelligence(AI) technologies. As we all know, the better data quality, the better performance of machine learning and deep learning models. How good is the data quality and how to control the data quality? Answering this question needs the knowledge of data quality standards, data quality assessment methods and data quality control methods. Currently, the comprehensive analysis of this problem is lacking. Scholars from MIT write a few research works about data quality assessment framework or methods [1-3]. These research confirmed data quality is a multi-dimensional concept and proposed the subjective and objective assessment methodology. There are many other methodologies of data quality assessment for different data types in various area[4]. However, in this big data era, there are challenges to apply the assessment methods in practice. Researchers have been developing various data augmentation approaches to improve the performance of machine learning or deep learning models because they know that the good training data quality is the key for good performance algorithms. However, most of them haven't do data quality assessment quantitatively before building the target models[5,6].

Controlling and improving products quality has become an important business strategy for many organizations: manufacturers, companies, transportation, financial services organizations, and government agencies. In manufacturing, the processes to produce data have some similarities to the processes that produce products. Data can be viewed as product which can be measured and the processes can be viewed as producing data products for data users and consumers. In the real world, statistical control methods are often implemented to the production to monitor the process and improve the product quality [7,8]. These statistical control process methods have achieved great success in automobiles, computers, energy, healthcare and other services[9-14]. For example, the control chart has been widely used in production line in manufacturing and six sigma management have been widely accepted by many organizations. This great success inspired us that the statistical process control methods could be used to control and improve the quality of data. To our best knowledge, there are little research publications on using the statistical control methods to control and improve the data quality in the data production process.

Data of high quality means the data are fit for use by the data consumers. Data consumers wants to get knowledge and value from the data by applying data mining approach, AI models or data fusion methods. A lot of research work show that machine learning and deep learning have significant potential to help solve challenging problems in many cases. In the past, the primary goal is to find the best models which can achieve the best performance. However, the best model achieved perfect result in the lab but impotence in practice. More recently, many companies are trying to improve the pre-processing procedure or training data quality for better model performance. The model-centric AI is developing to data-centric AI [15-19]. The key idea of data-centric AI is improving the data quality by

various tools. This trend is motivating researchers and engineers to investigate data deeper and improve data quality. For example, new data augmentation methods or training data selection methods are proposed for machine learning and deep learning models[20,21]. The data augmentation method increases the imbalanced level and generates more positive data to improve the supervised learning results. The training data selection method is developed to improve the performance of deep learning to predict the anomalies in multiple products production line. The multiple products production line is a data production process. The training data selection process can be viewed as a training data production process for specific deep learning models. To our best knowledge, there is almost no research work about using data quality control methods to monitor and improve the training data quality in the training data production process for deep learning models. In this study, the control chart is introduced and applied to the training data production process when building the anomaly prediction models.

The rest of the article is organized as follows. Section 2 presents the proposed data assessment metrics, the basics of control charts and improvement framework of data quality. Section 3 shows a real-world case study and explains the details of quality assessment and the design of training data quality control charts. Section 4 concludes the results and discusses the future research directions.

II. THE PROPOSED DATA QUALITY AND CONTROL METHOD

In this section, a subjective and objective data quality assessment metric is designed for imbalanced multivariate time series data. The basics of a control chart is introduced to be used in the data collecting or training data selection process. The framework of data quality assessment and control is proposed.

A. The basics of the data quality assessment method

Data quality is a multi-dimensional concept. We may define data quality in many ways according to the situation. Both subjective perceptions and the objective measurements should be deal with. Most people may have a conceptual understanding of quality as relating to some desirable characteristics. The individuals who assess data subjectively could be domain experts or data scientists. Although the conceptual understanding is a useful starting point, we prefer a more precise definition.

A research group from MIT defined the data quality dimensions and proposed objective assessment can be task-independent or task-dependent [1]. This means the quality of data can be described and evaluated by several dimensions. The task-independent metrics can be applied to any dataset to reflect the states of the data without the contextual knowledge applications. Task-dependent metrics are developed in specific application contexts which may include business rules, government regulations and constraints provided by the database administrator. They provide an excellent discussion and summarization of the dimensions of data quality. They described sixteen data quality dimensions and present three functional forms for developing data quality metrics objectively. The sixteen dimensions are accessibility, appropriate amount of data, believability, completeness, concise representation, consistent representation, ease of manipulation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, understandability and value-added. In practice, it is not easy to use the idea directly. As we all know, the data quality affects the performance of algorithms. To evaluate the data quality before building the models can help improve the understanding of the data and the choice of the algorithms.

In manufacturing industry, the multivariate times series data collected from the production process can be assessed by the user-defined data quality metrics. There are failures happens in the manufacturing process which should be prevented. This kind of multivariate time series data are often imbalanced. We add the imbalance level to the data quality metrics. We choose the typical dimensions and define the data quality assessment metric which can be used to evaluate the imbalanced data. The details of the definition are described in table 1. The imbalance level is defined as the ratio of abnormal data to normal data and the value is in [0,1]. For multivariate time series data, the final value of evaluation can be a weighted average of the variables. There are continuous variables and categorical variables in multivariate time series. The categorical variable is hard to be used to train a model. The ease of manipulation can be the ratio of number of continuous variables to the total number of continuous variables and categorical variables. Not every variable is relevant to the task or can be used for building the model. The value of relevancy can be decided by the domain expertise. The first step of using the data is checking whether the data is correct or not. The value of free-of-error reflects the reliability of the dataset. The parameters w_1 and w_2 are the customized weight and their values are fall in [0,1]. The weighted average is the final result of assessment.

TABLE I. DATA QUALITY METRICS FOR IMBALANCE DATA IN MANUFACTURING PROCESS

Dimensions	Definitions
Free-of-error(FE)	The extent to which data is correct and reliable
Appropriate Amount of Data(AAD)	The extent to which the volume of data is appropriate for the task
Ease of manipulation(EM)	The extent to which data is easy to manipulate and apply to different tasks.
Relevancy(R)	The extent to which data is applicable and helpful for the task
Imbalance level(IL)	The ratio of abnormal data to normal data
Weighted Average	$w_1 * (FE + AAD + EM + R) + w_2 * (IL)$

The first four dimensions in table 1 are chosen from the research work of MIT which we think are suit for assessment in manufacturing process. The last two dimensions are defined by ourselves. The dimension of imbalance level is a popular used concept in imbalanced learning. In the real-world manufacturing practice, there are little failure or anomaly data but big normal data about the production. The imbalanced data often creates bias in training processes of machine leaning or deep learning and sometimes make it not sufficient to support the decision-making systems. As more widely use of machine learning and deep leaning models in practice, the imbalanced level is an

important point to consider. In the weighted average equation, the imbalanced level should be given a bigger weight.

B. Data quality control chart in data production process

Control charts are widely used to establish and maintain statistical control of a process. The design of the control chart is to select the parameters required. Traditionally, the engineers or analyst select the parameters include a sample size, interval between samples and the control limits. There are different kinds of control charts for different situations. Shewhart chart, CUSUM chart and EWMA Chart are popular in industry practice [7]. Shewhart $\bar{X}$ chart is the first control chart developed by Walter A. Shewhart in 1925 based on the principle of 3σ. In this paper, the idea of this $\bar{X}$ control chart is designed to control data quality.

A typical control chart is shown in figure 1. It is a graphical display of a quality characteristic that has been computed or measured from a sample versus the sample time or number. A center line represents the average value of the quality characteristic. The upper control limit (UCL) and the lower control limit (LCL) are chosen to judge if the process is in control. As long as the points plot within the control limits, the process is assumed to be in control. However, a point that plots outside of the control limits is interpreted that the process is out of control.

Data is viewed as a product in this study and it can be measured by quality. From the start to the end of a process, the data flow. The control chart can also be used to measure the data quality in training data production process.

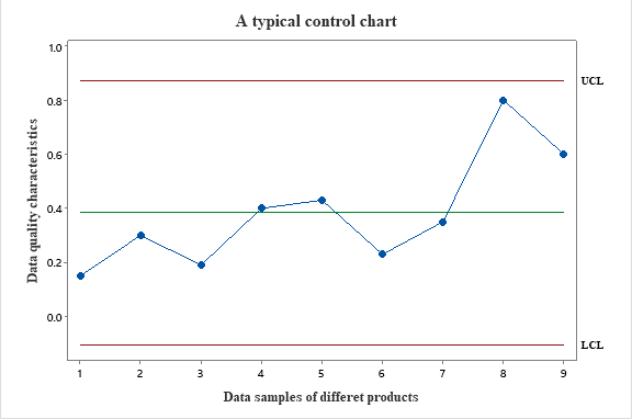


Fig. 1. A typical control chart to measure data quality

C. The proposed framework of data quality asessment and control

Nowadays, machine learning and deep learning methods are widely used to monitor or control the complex manufacturing process. The training data quality affects the training performance of the algorithm. To assess the data quality before training the algorithm helps us understand the data better and gives us ideas on how to improve the data quality. If the result of data quality assessment is not satisfied to the requirement, new data should be collected or the data quality should be improved by technical tools. The flowchart of data quality assessment and control framework is shown in figure 2.

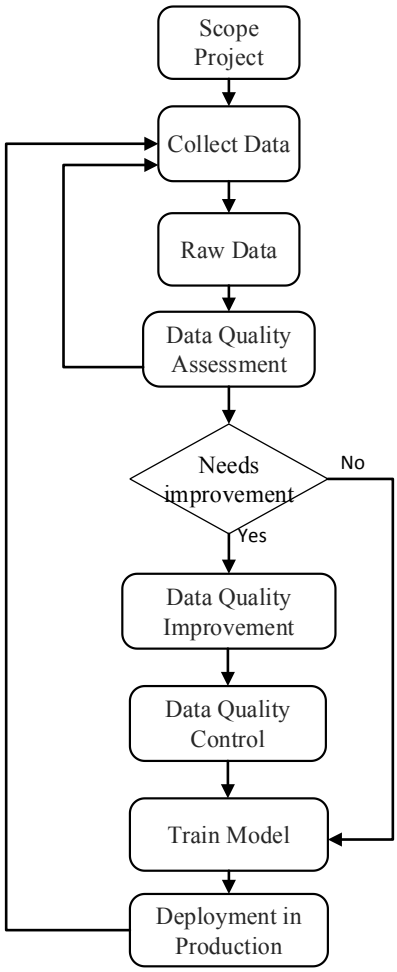


Fig. 2. The flowchart of data quality assessment and control framework

In the next section, a typical real-world multiple products manufacturing process is introduced and the proposed ideas are combined to improve the training data quality and the performance of the algorithm trained before.

III. REAL CASE STUDY

In this section, the paper manufacturing process is chosen as an example of a multiple products manufacturing process. The details of case data and result of the application are described.

Pulp and paper production is one of the largest manufacturing sectors in the world and it is a highly complex process by chemical and mechanical means. In one production line, many different grades of paper are produced. At the end of the production, paper breaks sometimes happen and they cause huge loss for the mills. There are multiple sensors located along the machines to collected data which can reflect the state of the production. The data collected can be viewed as imbalanced multivariate time series. In the earlier work about this case, deep learning and machine learning methods are used to detect and predict the failures [20,21]. The data is public and the link of the data can be found in the citation [22]. There are 59 continuous variables and 2 categorical variables. One categorical variable represents different paper grades. The amount of paper is decided by the market and customers. For different grades of

paper, there are often different length of multivariate time series. In practice, the production should produce new paper grades which may not be produced before. This situation brings challenges for the failure detection and prediction models because the models are trained by the historical data. For the new data of new paper grade, the trained models may not work well. For this situation, the data quality should be assessed and control before building the model.

According to the data quality metrics in table 1, the value of the assessment is shown in table 2. The subjective assessment is done by five different domain engineers and the imbalance level is equal to the ratio of positive data. All the values are between [0, 1]. The mean values of each dimension are calculated. For imbalanced data, the imbalance level is important and we give a bigger weight than other four dimensions. $w_1 = 0.35, w_2 = 0.65$. Different users may give different weights. We suggest the first weight should not bigger than 0.5. The value of weighted average is between [0, 1]. The result of the weighted average is the final assessment of data quality. Values close to 1 represent good quality while values close to 0 represent weak data quality. In this case, the final result is 0.2 which means the data quality should be improved.

TABLE II. RESULT OF THE SUBJECTIVE ASSESSMENT AND OBJECTIVE ASSESSMENT OF DATA QUALITY

Dimensions	**Value**					**Mean**
Free-of-error	0.9	0.8	0.9	0.8	0.9	0.86
Appropriate Amount of Data	0.6	0.7	0.6	0.6	0.7	0.64
Ease of Manipulation	0.3	0.2	0.4	0.3	0.4	0.32
relevancy	0.3	0.5	0.4	0.4	0.5	0.42
Imbalance level	0.0067					0.0067
Weighted average: 0.35*0.25*(0.86+0.64+0.32+0.42)+0.65*0.0067=0.2						

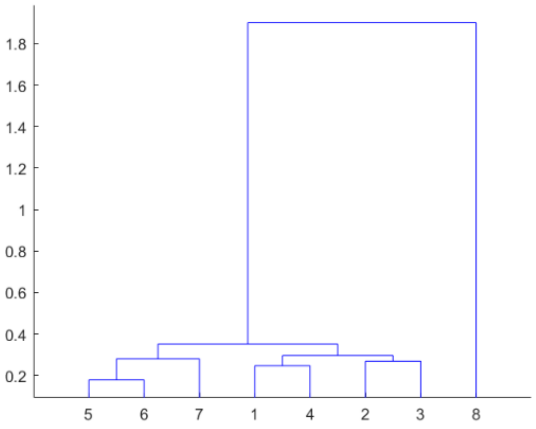
In our research work [21], a novel data augmentation is developed to improve the performance of failure detection models which use the supervised machine learning models logistic regression and random forest. This shows that the data quality assessment is meaningful for the model builders and industry.

In our research work [20], deep learning models have been used to predict the breaks and the training data is spilt depend on the result of the training data selected result. The training data selection method is composed by the feature similarity distance metric and hierarchical clustering of the distance. The similarity distance is shown in table 3 and the result of hierarchical clustering is shown fig 3(a) which the details can be checked in the citation [20].

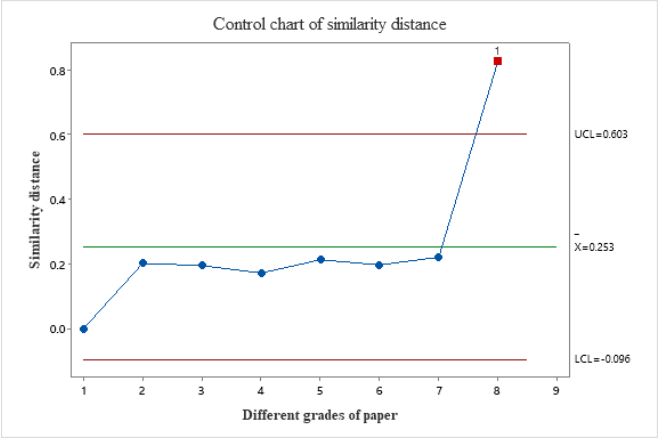
TABLE III. THE SIMILARITY DISTANCE OF EACH PAPER GRADE FROM THE DISTANCE MATRIX[20]

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Paper Grade	**96**	**82**	**118**	**139**	**112**	**84**	**93**	**51**
Distance	0	0.201	0.195	0.172	0.213	0.197	0.221	0.827

The details of how to calculate the distance which represent the similarity of different samples can refer to our research work [20]. In this study, we directly use the results of distance for training data selection. For the first cluster, the mean of distance is the center line. Because the element of distance matrix starts from zero, the value of the control limit is moved up for better plot. We set the lower control limit and upper control limit. The control chart is plot in figure 3(b).



(a).



(b).

Fig. 3. The control chart for the training selection method

According to the clustering results in fig 3(a), we know that there are different groups of paper grades. In the research work [20], we divide them into 2 groups. However, if only based on the clustering result, we may have three or four groups. For the new coming data samples of new paper grades, we may have no ideas about how many groups should be redivided. However, for the control chart, it shows more clearly the distance of different data samples and threshold of the group. For any other grades of paper, if the distance is out of the control limit, the control chart

will generate alarm. It means the data should be selected for the other cluster and fix to another autoencoder model.

This control chart can not only be used to monitor the similarity of different grades of paper but also can generate the alarms which can used to remind the engineers the unsatisfied data sample quality of new paper grades for the deep learning model.

IV. Conclusion and Discussion

In this study, a data quality assessment metric is designed for imbalanced multivariate time series data from multiple products manufacturing process. The result of the assessment can reflect the quality of imbalanced data should be improved. This study combined the valuable result of previous research work to show the quantitatively assessment of data quality is part of the data-centric AI. We proposed the idea that the training data production process could be controlled like a product production process. The data quality control chart is designed and the result show that it is better than the previous clustering result. More importantly, the data quality control chart works for the new coming data samples and it can significantly detect the new pro In training data production process, it can generate the alarms when it is out of control online and remind the domain experts. These results help the domain experts understand the data deeper and help do the right training data selection for deep learning models to get better performance.

Data can be viewed as a product. The methodology of data quality assessment is developing and the trend of its development is from qualitatively evaluation to quantitatively evaluation. In practice, data quality assessment is not an easy task. The dimensions of the data quality could be chosen from the previous work or designed depend on the tasks. The dimensions should be precisely defined first and the value of each dimension should be given by domain experts or calculated by equations. There is no best data quality metric.

The idea of data production process can be controlled proposed in this study is new. There are many potential future directions of this topic such as define more precise data quality assessment dimensions for single or multiple signals in manufacturing, define standard data quality assessment metrics for data collected from one production line, design data quality improvement cycle.

Acknowledgment

This research is supported by National Natural Science Foundation of China (Grant No.51775108 & No.51705238).

References

[1] Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211-218

[2] Wang, R. Y., Ziad, M., & Lee, Y. W. (2006). *Data quality* (Vol. 23). Springer Science & Business Media.

[3] Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE transactions on knowledge and data engineering*, *7*(4), 623-640.

[4] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 1-52.

[5] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.

[6] Lee, Y. W., & Strong, D. M. (2003). Knowing-why about data processes and data quality. *Journal of Management Information Systems*, *20*(3), 13-39.

[7] Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

[8] Tsung, F., Li, Y., & Jin, M. (2008). Statistical process control for multistage manufacturing and service operations: a review and some extensions. International Journal of Services Operations and Informatics, 3(2), 191-204.

[9] Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: an overview. Quality and Reliability engineering international, 23(5), 517-543.

[10] MacCarthy, B. L., & Wasusri, T. (2002). A review of non-standard applications of statistical process control (SPC) charts. International Journal of Quality & Reliability Management.

[11] Baldassarre, M. T., Caivano, D., Kitchenham, B., & Visaggio, G. (2007, April). Systematic review of statistical process control: An experience report. In 11th International Conference on Evaluation and Assessment in Software Engineering (EASE) 11 (pp. 1-9).

[12] Benneyan, J. C., Lloyd, R. C., & Plsek, P. E. (2003). Statistical process control as a tool for research and healthcare improvement. BMJ Quality & Safety, 12(6), 458-464.

[13] Castillo, E. D., Grayson, J. M., Montgomery, D. C., & Runger, G. C. (1996). A review of statistical process control techniques for short run manufacturing systems. Communications in Statistics--Theory and Methods, 25(11), 2723-2737.

[14] Peres, F. A. P., & Fogliatto, F. S. (2018). Variable selection methods in multivariate statistical process control: A systematic literature review. Computers & Industrial Engineering, 115, 603-619.

[15] Alvarez-Coello, D., Wilms, D., Bekan, A., & Gómez, J. M. (2021). Towards a data-centric architecture in the automotive industry. Procedia Computer Science, 181, 658-663.

[16] Whang, S. E., Roh, Y., Song, H., & Lee, J. G. (2021). Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *arXiv preprint arXiv:2112.06409*.

[17] Bartel, C. J. (2021). Data-centric approach to improve machine learning models for inorganic materials. Patterns, 2(11), 100382.

[18] Abedjan, Z. (2022). Enabling data-centric AI through data quality management and data literacy. it-Information Technology, 64(1-2), 67-70.

[19] Motamedi, M., Sakharnykh, N., & Kaldewey, T. (2021). A data-centric approach for training deep neural networks with less data. arXiv preprint arXiv:2110.03613.

[20] Xu, D., Zhang, Z., & Shi, J. (2022). Training Data Selection by Categorical Variables for Better Rare Event Prediction in Multiple Products Production Line. *Electronics*, *11*(7), 1056.

[21] Xu, D., Zhang, Z., & Shi, J. (2022). A New Multi-Sensor Stream Data Augmentation Method for Imbalanced Learning in Complex Manufacturing Process. *Sensors*, *22*(11), 4042.

[22] Chitta Ranjan. Understanding Deep Learning: Application in Rare Event Prediction.1st ed. Connaissance Publishing;2020.