Co-evolution of Measurement Oriented Information Processing Technology and Information Processing Oriented Measurement Technology.

Takashi WASHIO

Professor
Department of Reasoning for Intelligence
Division of Information and Quantum Sciences
The Institute of Scientific and Industrial Research
Osaka University
Doctor of Engineering

With the advent of the IoT society, there is an increasing need for measurement techniques that use the principle of complex processes under extreme conditions. On the other hand, rapidly developing information processing using machine learning and statistical estimation is good at making highly accurate and reliable estimation from incomplete and complex large amounts of information. From such a background, the systematic co-evolution of measurement-oriented information processing and information processing-oriented measurement technology is opening up the possibility of realizing new advanced measurement technology that meets the needs of the IoT society. This paper describes some of the research issues that have become clear in the field of measurement-oriented information processing “measurement informatics”, and further introduces the outcomes of utilizing that principle in the development of advanced information processing-oriented measurement devices and equipment.

Introduction

Measurement technology that digitizes and collects necessary information from the real world, along with information communication technology and information processing technology, is a core technology required for the future of the IoT society. In the new IoT era, in addition to measurement technology with conventional sensors, there is a great need for so-called advanced measurement technology including measurement technology that measures completely new objects as well as measurement technology that enables robust, compact and low cost sensing with more frequency, higher accuracy, higher resolution, and higher reliability.

Based on these needs, research and development of various advanced measurement devices and equipment is currently being conducted. There is a wide range of measurement techniques, such as extremely low concentration substances, extremely small molecules and organisms, images and material distributions in ultra-high resolution, very distant objects, ultra-high frequency / short time extremes, and ultimate compactness of size. Many aim at measurement and information processing under these extreme conditions, and use the measurement principle and calculation principle consisting of complicated processes under extreme conditions. Furthermore, the object to be measured is often a complicated one that contains large fluctuations in both time and space, and has a very high degree of freedom. Therefore, the obtained measurement information often contains a large amount of noise and only incomplete information on the measurement object. Consequently, advanced estimation processing is required to obtain sufficient measurement results, such as target estimation from incomplete large amounts of information, integration of multiple measurement information, and interpolation using prior knowledge.

In parallel, advanced mathematical and statistical theories, and machine learning / statistical estimation based on computational theory are showing rapid development. They are good at making estimations with high accuracy and reliability from the above-mentioned incomplete large amounts of information with multiple measurements and prior knowledge. From such a background, by combining machine learning / statistical estimation / algorithms with advanced measurement devices and equipment, the possibility of realizing new advanced measurement technology that meets the needs of the IoT society is about to widen.[1]

However, to date, such a fusion of research and

development has only been carried out sporadically and on individual problems. There has been very little systematic research on the principle of machine learning, statistical estimation, and algorithms directed towards advanced measurement as well as new measurement technology directed towards realization of highly efficient measurement based on these information processing techniques. In order to meet the strict requirements for measurement in the IoT society, systematic co-evolution of measurement-oriented information processing and information-oriented measurement technology is required.

The author is engaged in the development of "measurement informatics", which is a research field of machine learning, statistical estimation, and algorithm principles that directs measurement from the viewpoint of information processing research. We collaborate with many researchers and engineers who are engaged in research and development of cutting-edge measurement devices and equipment and assist in the development of measurement technology directed towards information processing. This paper describes some of the issues in information processing research that are becoming clear in the field of measurement informatics research and further introduces some of the developments of information processing-oriented, advanced measurement devices and equipment that makes use of the principles arising from that research.

Current Status and Research Issues on Measurement Informatics

Measurement informatics research that has undergone recent rapid development on the principles of machine learning, statistical estimation, and algorithms has just begun. Even overseas, only sporadic research and development on individual issues has been conducted, and there has been no systematic research activity organized. One of the causes is that most artificial intelligence research including machine learning, statistical estimation, and algorithms pursue versatility widely applicable to all fields including various services, finance, natural sciences, and engineering. Although its broad applicability is a great source of social impact of artificial intelligence and should be evaluated, problem establishment inherent in use in specific fields of measurement is not considered, resulting in a "poor fit" from the basic research stage. Another cause is that there are few researchers and engineers who are familiar with both advanced information processing and advanced measurement devices while data scientists worldwide are in very short supply. In addition, the current environment growing such human resources is still premature.

As described above, although still in its infancy, some approaches that require systematic thinking at the basic research level have become increasingly clear in applying information processing technology to measurement through the study of measurement informatics. Below, we will explain three main issues found in our research, and it is expected that many more issues will be found as the field continues to develop.

Point 1: Estimation for analysis and estimation for measurement[1]

Machine learning and statistical estimation are theories and techniques that reveal some regularity found in given individual dataset. These are intended to derive the regularity embedded even in a biased dataset without considering natural laws and common sense in the real world. While this property allows good estimation for the domain of the given dataset, it may give false results for problem settings that deviate from it.

For example, you might have seen an image of a dog or a wolf a few dozen meters away in the city of Kyoto. It is reasonable to assume that it is a dog because there is little possibility of a wolf in Kyoto. In standard machine learning and statistical estimation, if a lot of data in which photographs of the image of a dog or a wolf seen from a distance of several tens of meters in the city of Kyoto are collected and learned with high accuracy by a learning algorithm, the resultant algorithm presumes that the image is a dog, even if the image which closely resembles a wolf is seen. The reason is that there are few images of wolves in the data for learning, and the data properties are included in the learnt algorithm that is almost always correct when it is presumed to be a dog. In other words, not only measurement results based on individual images but also those by additionally taking into account the regularity indicated by the entire data showing that wolves were infrequently seen are obtained. This is similar to the fact that a doctor backed by years of experience takes into account the past findings of the doctor in addition to the nature of the patient's X-ray image, and guides the diagnosis result accordingly. If the problem to be solved is an analysis that adds the experience backed by past data to the results obtained from the measurement equipment and devices, standard machine learning and statistical estimation techniques may be applied.

On the other hand, if we take the estimation equipment as is, using the data collected and learned in Kyoto into the Himalayan forest for example, we continue to assume that it is a dog despite the fact that most of the images of animals, which resemble either dogs or wolves, are in fact wolves. This is a problem in measurement methodology. In the measurement, even if the analysis based on past experience is not useful, it is required to always output

the correct estimation regardless of the environment. For this purpose, it is necessary to use an algorithm that predicts a dog or a wolf based solely on the features of individual images, without using the nature or regularity of the entire learning data. In order to accomplish this by means of existing machine learning and statistical estimation techniques, it is necessary to devise a method such as using learning data consisting of a greater number of balanced dog and wolf images. However, if the information you want to estimate is not a simple choice like dogs and wolves, but complex images and spectra, it is infeasible to prepare learning data without bias including all possible unknown images and spectra. Therefore, for many complex advanced measurements, research and development of new learning and estimation principles directed towards measurement will be required.

The estimation for the analysis and the estimation for the measurement described above have long been called "Maximum A-Posteriori Probability (MAP) estimation" that is based on the distribution of the data, and "Maximum likelihood (ML) estimation" that does not refer to the distribution of the data, respectively. Many of the machine learning, statistical estimation principles and algorithms currently being researched and developed, including deep learning, are the former, and it is often necessary to rework the basic research level for measurements. Research on machine learning and statistical estimation based on ML estimation aiming at measurement is an important issue of measurement informatics.

Point 2: Bayesian estimation[2, 3]

In measurement and analysis, in addition to the regularity indicated by the distribution of data described above, it may be desirable to make an estimation that reflects the nature of the estimation target that we know in advance. For example, if it is known in advance that the distribution of the protein to be measured is localized in a part of the cell and does not spread in the entire measurement image, it is possible to perform more appropriate protein distribution estimations, reducing the influence of observational noise. An estimation that reflects such prior knowledge is called "Bayesian estimation".

Especially, such estimation that reflects knowledge that an object should be localized to some parts of measurement results or analysis results, as in the above example, is called "Sparse estimation" and is used in many machine learning and statistical estimations. Also, another Bayesian estimation method has been proposed that reflects prior knowledge that an object should be present with a smooth distribution, or in several separate chunks of the analysis.

However, our diverse prior knowledge cannot be easily introduced into estimations. In order to reflect prior knowledge in estimation, it is necessary to formulate knowledge mathematically, and also to find an algorithm that enables easy calculation of the formula. There is no guarantee that any prior knowledge formalization or its efficient calculation algorithm exists, and if it does indeed exist, experts in the field of mathematics and algorithms are required to find them. The clarification and standardization of our prior knowledge and the method to introduce such knowledge into Bayesian estimations widely and easily is an important issue in measurement informatics.

Point 3: Sensor arrangement[4, 5, 6]

Many advanced measurements use measurement principles that consist of complex processes under extreme conditions. In particular, many measurement and analysis techniques use multiple sensing devices in combination to obtain good results. In many of these cases, the number of sensors and their arrangement are very important factors to maximize the accuracy and robustness of measurement and analysis results under resource constraints such as cost and device configuration.

It is known that many of the problems to find an appropriate number of sensors and their arrangement can be mathematically formulated as the maximization problem of submodular functions. For example, consider the problem of how many surveillance cameras can be used and how they are arranged in a building to construct a system for monitoring the area in the building more widely and effectively. The monitoring area in this case is determined by selecting which, and how many locations are to be used from all the potential locations where cameras can be installed. The monitoring area is a function of an installation point set, and is called a set function. The area that can be monitored with only one camera is limited, and the monitoring area can be expanded by optimally arranging additional cameras. However, if the number of cameras is increased by too much, the coverage will overlap and the monitoring area will not be significantly increased. In other words, the monitoring area increases only gradually relative to the number of cameras installed, following the principle known as the "law of diminishing returns". In general, a set function that obeys the law of diminishing returns is called a "submodular function". The problem of finding a set of sensor locations that maximizes the output of a submodular function such as the monitored area is called the "submodular function maximization problem".

In many measurement and analysis problems, various performance indicators such as accuracy and robustness

are known as submodular functions. Therefore, it is important to solve the submodular function maximization problem in this field. Upon determining the optimum sensor arrangement, however, it is necessary to try a large number of combinations from a large number of installation candidate points in order to select the optimum installation point set that maximizes the monitoring area. In the field of computational science, finding a set that strictly maximizes a submodular function is known to be the "NP hard problem" which is a very computationally intensive problem. Algorithms that solve large-scale problems in a practical speed with high accuracy are still under study. The development of such practical algorithms is also an important issue in measurement informatics.

Research and Development Cases

Here, two research cases that were conducted jointly by measurement technology researchers and the authors, as information processing researchers, are described. The two cases are related to the advanced measurement technology to which, among the above-mentioned measurement informatics, Point 1 -estimation for analysis and estimation for measurement- is applied.

Case 1: Analysis of ultrafine particle type using nanopores[7-10]

This case is an authors' joint study with Prof. Masaki Taniguchi and his colleagues who are members of The Institute of Scientific and Industrial Research, Osaka University to which the author belongs. They work on nano measurement technology research at the Department of Biotechnology of the Division of Nanoscience and Nanotechnology Center.

The nanopore is a nano-scale hole filled with an electrolyte solvent as shown in Figure 1. When a constant voltage is applied by arranging the electrodes above and below, the ion current flows through the nanopores. When the ultrafine particles are introduced from the upper side by the action of an electric field or pressure and passed through the nanopores toward the lower side, the ion current is lowered because the particles temporarily partially block the hole during passage, and a reduced current pulse is obtained. This pulse shape reflects not only the size and the velocity of passing ultrafine particles but also various other information such as their shape and surface condition. Therefore, it is possible to analyze the type and property of the ultrafine particles which pass individually by the pulse shape output from the nanopore.

In the ultrafine particle identification by the conventional nanopore output pulse, as shown on the left side of Figure 1, the identification was performed based on the wave height and the wave width of each pulse. The larger the particle size, the larger the clogged area of the nanopore, so it was

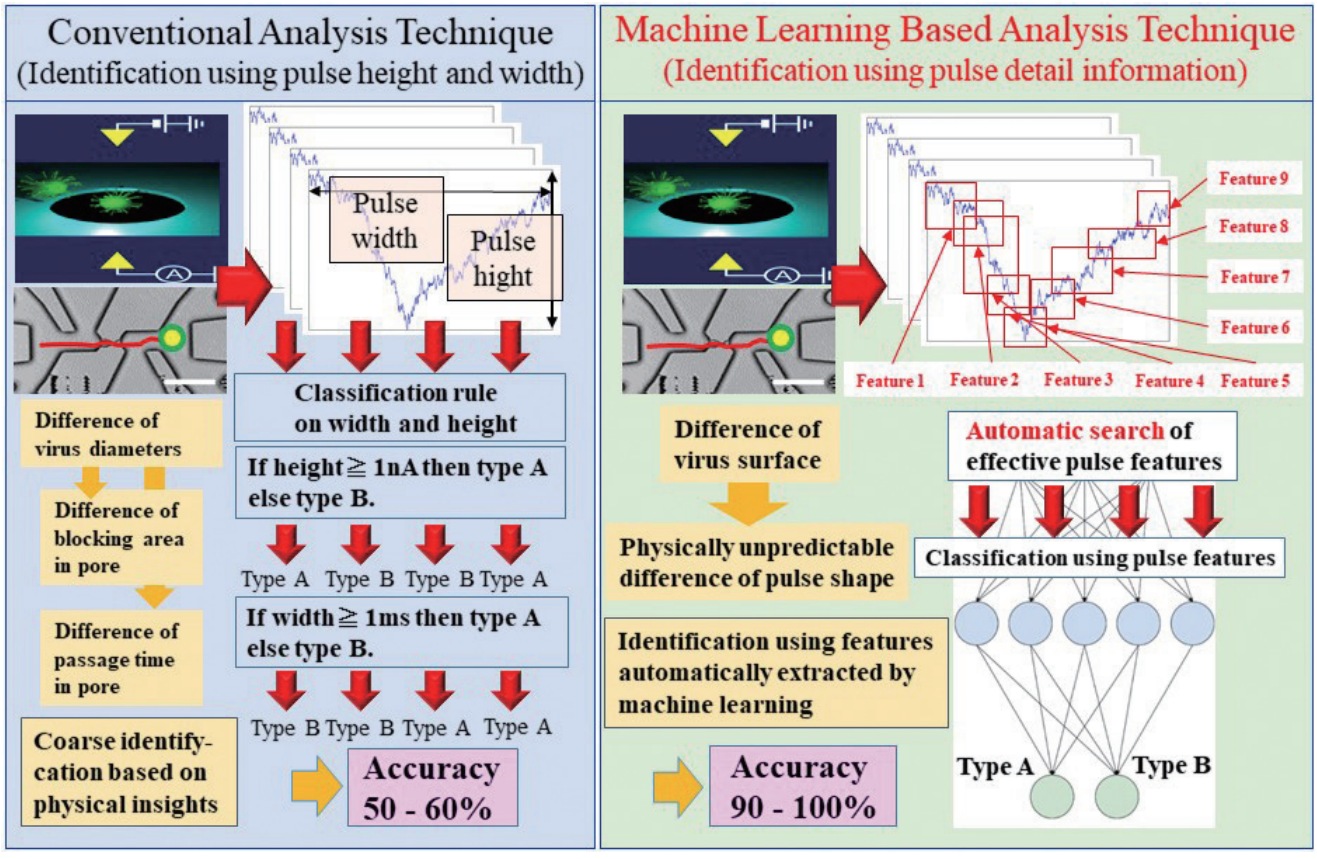

Figure 1 Difference between conventional nanopore pulse identification and nanopore identification using machine learning.

assumed that the pulse wave height would be larger. If the particle has a size, mass and surface condition that easily passes through the nanopore, it is also assumed that the pulse width would be small because the particle passes through the nanopore at higher speed. Conventionally, based on such physical knowledge, attempts have been made to identify particle types by determining the threshold of wave height and wave width of the pulse. However, even with the same type of particle, there are individual differences in size and state, and the difference in type is not necessarily reflected clearly in wave height and wave width. Furthermore, the current pulse to be measured also has a lot of noise. Therefore, except in the case of distinguishing particles with significantly different sizes and properties, it is often the case that sufficient identification accuracy cannot be obtained by relying only on the wave height and wave width of the pulse.

In this joint research, as shown on the right side of Figure 1, the features of the pulse waveform are indexed as feature quantities. As feature quantities, in addition to the conventional wave height and wave width, coarse grained pulse wave height with a strip-like rough time interval, various indicators showing the degree of peaking and deviation of pulses and indicators automatically generated by machine learning technology are used. A pulse waveform dataset was acquired for each type of ultrafine particle to be identified, and machine learning algorithms were applied to this dataset to obtain classifiers that identifies the type of particle. The machine learning method incorporates an algorithm that automatically selects the feature that is effective in identifying the target particle type so that robust identification can be performed against individual differences in particles and measurement noise.

Here, the learning data is prepared based on the frequency at which various ultrafine particles appear in the sample, so that the obtained classifiers reflect not only the pulse waveform characteristics but also the frequency of appearance of particle types when analysing the data distribution. In other words, identification of particle types based on the analysis considering actual data distribution is performed.

This technique was applied to the identification of pairs of bacteria with similar properties and shapes, such as E. coli and B. subtilis, S. epidermidis and S. aureus. While conventional methods could only obtain 50 to 60% identification accuracy, the method used here achieved 90% - 100% identification accuracy. When applied to identification of 3 types of influenza virus, type A, type B and subtype A, conventional methods achieved an accuracy of 50% or less, while this method obtained a 72% level of accuracy. Although the accuracy is not sufficient for each pulse, an algorithm that discriminates based on multiple pulse measurements was built, and a discrimination accuracy of over 99% was achieved.

Case 2: Ultra-robust olfactory measurement by MSS sensor[11, 12]

This case is an authors' joint study with Prof. Genki Yoshikawa and his colleagues who are engaged in olfactory sensor technology research in the nanomechanical sensing group, Nano-System Field, International Center for Materials Nanoarchitectonics, National Institute for Material Science.

The MSS olfactory sensor consists of multiple distinct MSS sensors as shown in Figure 2. Each sensor is a device that senses the stress generated by the thin film that is dynamically expanded and contracted due to adsorption and desorption of a trace amount of gas compound molecules contained in the air, and it converts the stress to a voltage signal. Therefore, even when exposed to the same gas compound molecules following the same concentration change in the air, different voltage waveforms are output from sensors consisting of thin films having different properties. The MSS olfactory sensor is composed of MSS sensors that have different response characteristics by mounting thin films of different properties, and based on the combination pattern of these output voltage waveforms, trace amounts of gas compound molecules that is the origin of odor in the air are identified. The sensor unit of the MSS olfactory sensor is made by MEMS technology and is compacted to several millimeters square in total size.

In the conventional olfactory sensor, as shown in the upper part of Figure 2, a mass flow controller or a pump is placed at the inlet, and the input air flow rate and the concentration of gas compound molecules in it are controlled in a rectangular wave, and each sensor response voltage waveform to the repetitive rectangular wave input is measured. Since the input waveform is completely controlled, the output voltage waveform of each sensor represents the characteristics of adsorption and desorption of each thin film on gas compound molecules. As this combination differs depending on the type of gas compound molecule, it identifies (learns) a classifier that performs maximum posterior probability (MAP) estimation by a conventional machine learning algorithm using previously collected data. Given a measured voltage waveform of an unknown gas compound molecule, its type can be estimated by this classifier. This is a classifier that makes the estimation for the analysis described in Point 1 and is trained to give the highest accuracy for the voltage waveform combination output for the square wave input. For the input other than the square wave input, the estimation

accuracy would be significantly degraded because the voltage waveform and the combination distribution therein are completely different, similar to the case where a classifier trained on the data obtained in Kyoto city is useless if it is brought to the Himalayan forest. Thus, the flow control of the input is required and the sensing equipment becomes much larger despite the fact that the sensor unit itself is as compact as several millimeters square.

On the other hand, in this joint research, as shown in the lower part of Figure 2, a machine learning algorithm was developed to obtain a newly developed classifier for maximum likelihood (ML) estimation. The classifier, which estimates the type of gas compound molecules without being affected by the input airflow rate or the waveform of the concentration of the gas compound molecules in the mixture, was developed by learning the features that depends only on the type of gas compound molecules, independent of the input wave form, from the voltage waveform combination output from the MSS olfactory sensor. This is a classifier that performs estimation for the measurement described in Point 1, and eliminates the need for an input flow control by a mass flow controller or a pump. As a result, a sensor that detects odor using only the MSS olfactory sensor main body of about several millimeters square in size was developed.

The MSS olfactory sensor was hand-held and given a light shake over the top of a beaker containing a single gas compound such as ethanol, water, heptane, and ethyl acetate, and the type could be identified with an accuracy of 99.6%. The same measurement was performed on beakers containing aromatic mixed compounds such as rosemary, red chili pepper and garlic, with an identification accuracy of 89% achieved. However, when the identification was tried by the conventional machine learning using only the MSS olfactory sensor main body, only very low accuracy, which is insufficient for practical use, was obtained. Thus, the development of machine learning algorithms suitable for measurement problems has enabled the realization of ultra-compact and ultra-robust measurements that cannot be achieved with the development of devices and equipment.

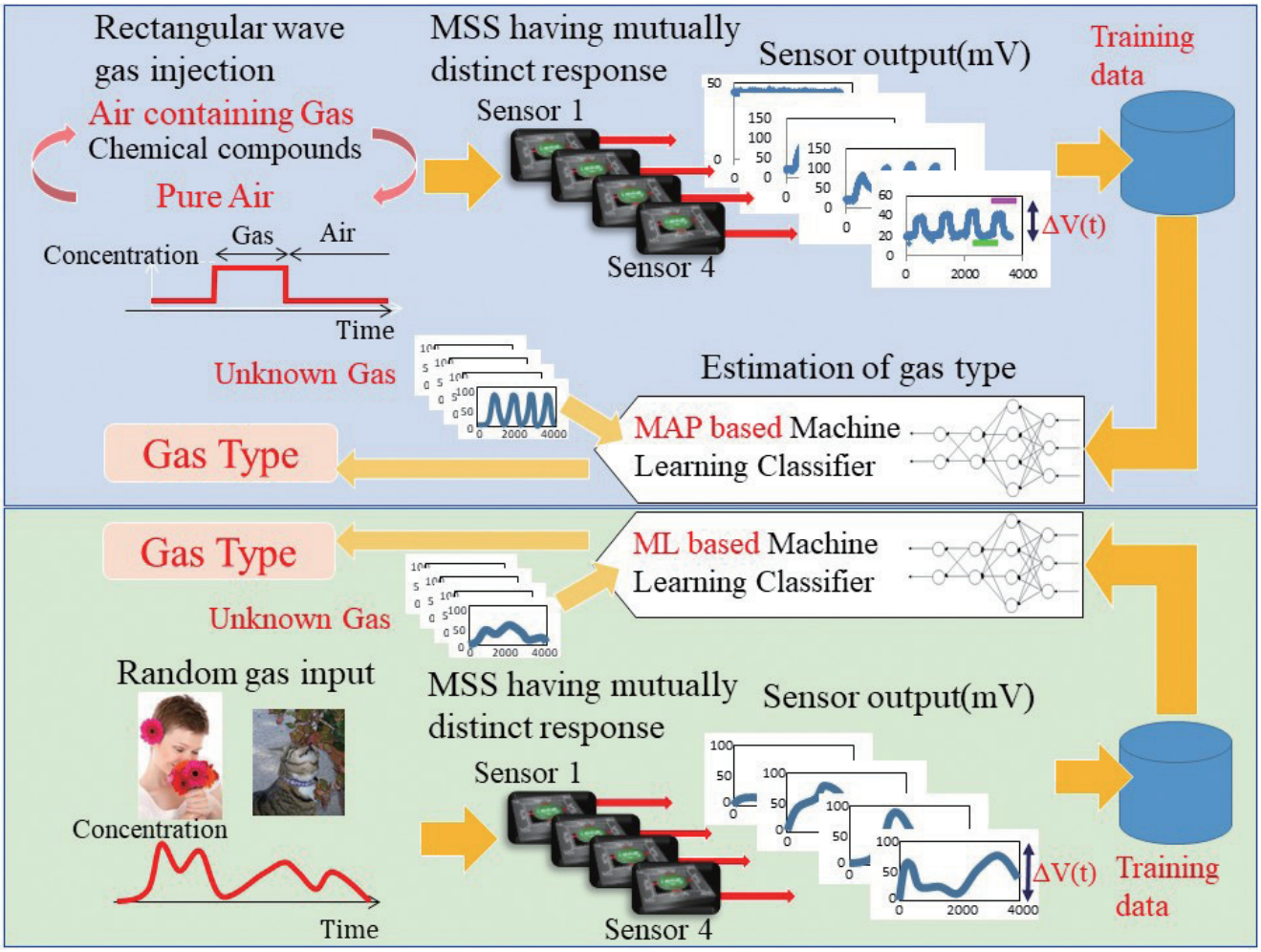

Figure 2 Difference between olfactory sensor by conventional machine learning and olfactory sensor by novel machine learning performing maximum likelihood estimation.

Conclusion

Along the development of the IoT society, the demand for measurement technology will become more and more advanced, and the advanced measurement will be required to combine more complex information processing and measurement device / equipment technology. For this purpose, systematic basic research and development on measurement and information processing is important. In this paper, measurement informatics aimed at systematically studying measurement-oriented information processing and some important issues therein were described, and examples of research and development of advanced measurement technology which make use of some of the outcomes were introduced. Along with advances in measurement informatics research, it is expected to see more results from a variety of measurement fields together with additional problems that will need to be addressed in the future.

The author aims to promote measurement informatics research, which is currently lagging behind in the world, to share information between researchers and engineers, and to develop human resources. In 2018 under the Japanese Society for Artificial Intelligence (JSAI), "Special Interest Group on Measurement Informatics: SIG-MEI"[13] was established, and research workshops are held twice a year. Even non-members of JSAI can participate and observe in the workshop. We expect your join into this group if you are interested.

References

[1] T. Washio, G. Imamura and G. Yoshikawa. "Machine learning independent of population distributions for measurement". Proc. DSAA2017: 4th IEEE International Conference on Data Science and Advanced Analytics, Tokyo (2017), DOI: 10.1109/DSAA.2017.28.

[2] Y. Nakanishi-Ohno, T. Obuchi, M. Okada and Y. Kabashima. "Sparse approximation based on a random over complete basis". J. Statistical Mechanics: Theory and Experiment (2016) 063302.

[3] T. Obuchi, Y. Nakanishi-Ohno, M. Okada and Y. Kabashima. "Statistical mechanical analysis of sparse linear regression as a variable selection problem". J. Statistical Mechanics: Theory and Experiment (2018) 103401, 1-41.

[4] Y. Kawahara, K. Nagano, K. Tsuda and J. Bilmes. "Submodularity cuts and applications". Advances in Neural Information Processing Systems (2009) 22, 916-924 (Proc. NIPS2009).

[5] Y. Kawahara and T. Washio. "Prismatic algorithm for discrete D.C. programming problem". Advances in Neural Information Processing Systems (2011) 24, 2106-2114 (Proc. NIPS2011).

[6] Y. Kawahara and K. Nagano. "Submodular Optimization and Machine Learning". Machine Learning Professional Series, Kodansha Scientific (2015).

[7] M. Tsutsui, Y. He, K. Yokota, A. Arima, S. Hongo, M. Taniguchi, T. Washio and T. Kawai. "Particle trajectory-dependent ionic current blockade in low-aspect-ratio pores". ACS Nano, American Chemical Society (2016) 10[1], 803-809.

[8] M. Tsutsui, T. Yoshida, K. Yokota, H. Yasaki, T. Yasui, A. Arima, W. Tonomura, K. Nagashima, T. Yanagida, N. Kaji, M. Taniguchi, T. Washio, Y. Baba and T. Kawai. "Discriminating single-bacterial shape using low-aspect-ratio pores". Scientific Reports (2017) 7(1) 17371.

[9] M. Tsutsui, M. Tanaka, T. Marui, K. Yokota, T. Yoshida, A. Arima, W. Tonomura, M. Taniguchi, T. Washio, M. Okochi and T. Kawai. "Identification of individual bacterial cells through the intermolecular interactions with peptide-functionalized solid-state pores". Analitical Chemistry (2018) 90, 1511-1515.

[10] A. Arima, M. Tsutsui, I. H. Harlisa, T. Yoshida, M. Tanaka, K. Yokota, W. Tonomura, M. Taniguchi, M. Okochi, T. Washio and T. Kawai. "Selective detections of single viruses using solid-state nanopores". Scientific Reports, (2018) 8, 16305.

[11] G. Imamura, G. Yoshikawa and T. Washio. "Development of machine learning models for gas identification based on transfer functions". Proc. 17th International Meeting on Chemical Sensors, Vienna (2018) AR1.1.

[12] G. Imamura, K. Shiba, G. Yoshikawa and T. Washio. "Free-hand gas identification based on transfer function ratios without gas flow control". Scientific Reports (2019) 9, 9768.

[13] http://www.ar.sanken.osaka-u.ac.jp/SIG-MEI/