The Most Advanced Data Mining of the Big Data Era

FUJIMAKI Ryohei, MORINAGA Satoshi

Abstract

Recently, the acquisition of knowledge from big data analysis is becoming an essential feature of business efficiency. However, the analysis of big data can be troublesome because it often involves the collection and storage of mixed data based on different patterns or rules (heterogeneous mixture data). This has made the heterogeneous mixture property of data a very important issue. This paper introduces "heterogeneous mixture learning," which is the most advanced heterogeneous mixture data analysis technology developed by NEC, together with details of some actual applications. The possibility of the utilization of data that has previously been collected without any specific aim is also discussed.

Keywords

big data, heterogeneous mixture data, data mining
machine learning, heterogeneous mixture learning, factorized asymptotic Bayesian inference

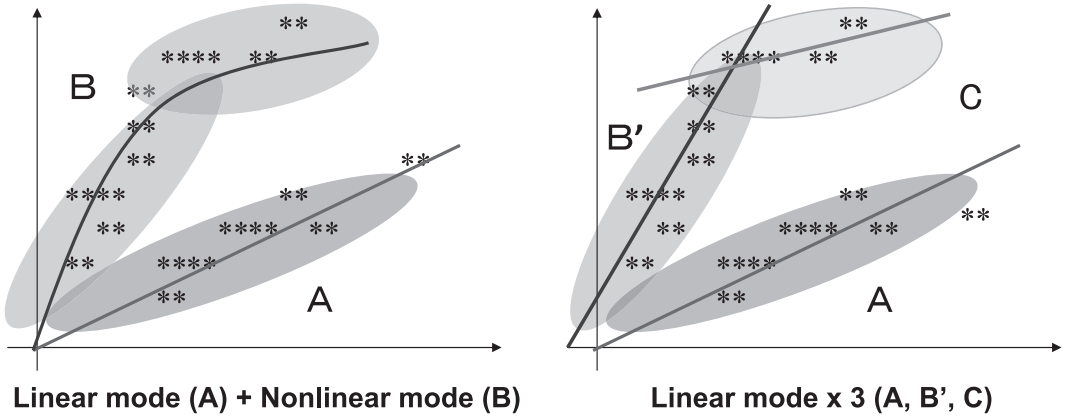
1. Introduction

Recent business trends suggest that big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently-occurring patterns and hidden rules. These may then be used effectively as valuable information [*1] (such knowledge-discovering technology is generally referred to as data mining). For example, electricity demand is predicted by extracting the rules governing the values of various sensors such as thermometers and of electricity demand and deriving future demand predictions by applying such rules to the current sensor data.

The difficulties of big data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data (heterogeneous mixture data issue). Especially, in the case of complicated heterogeneous mixture data, the data has not only several patterns and rules but characteristically, the properties of the patterns vary greatly (as shown in the left hand graph of **Fig. 1** , which shows a type of heterogeneous mixture data containing linear and nonlinear properties). In the case of electricity demand prediction, when the relationship between the sensor values and electricity demand are changed due to a specific cause, the conventional analysis technology has often been unable to determine the situation that led to degradation of the prediction accuracy. The countermeasure most often taken in such a case was to define factors altering the rules, such as the day of the week or time zone, by trial and error based on expert knowledge, and to classify the data accordingly in order to auto-extract individual rules. However, locating these factors is very difficult even for experts, which has led to issues such as the impossibility of defining heterogeneous mixtures due to insufficient data grouping or to the fragmentation of patterns caused by excessive data grouping (both of these issues could become the main cause of a drop in the prediction accuracy).

In this paper, we first discuss the difficulties of heterogeneous mixture data analysis. In short, the impossibility of performing exhaustive searches due to the huge number of data grouping candidates, which in reality symbolizes the essential difficulty of the analysis. Next, we introduce heterogeneous mixture learning. This is the most advanced heterogeneous data analysis technology to be developed at NEC. It features the application of an advanced machine learning technology called the "factorized asymptotic Bayesian inference," and we will focus mainly on the introduction of its fundamental concept. Finally, we introduce a demonstration experiment of electricity demand prediction for a building as an example of a suitable application of heterogeneous mixture learning. With the heterogeneous mixture learning technology, we have succeeded in improving the prediction

[*1] According to the 2012 survey [1)] made by the Yano Research Institute, Ltd., the big data market scale in FY2011 was 190 billion yen and that for FY2020 is expected to exceed one trillion yen.



For the proper analysis of data, it is required to find the optimum grouping method from a large set of data grouping candidates. (The ellipses correspond to the data grouping methods and the lines to the prediction models.)

Fig. 1 Illustration of heterogeneous mixture data.

accuracy by 7.6 points (10.3% → 2.7%) compared to the previous prediction method without considering the heterogeneous mixture data, and by 2.1 points (4.8% → 2.7%) compared to the method that is dependent on data grouping by experts.

2. Issues of Heterogeneous Mixture Data Analysis

One of the key points in the accurate analysis of heterogeneous mixture data is to break up the inherent heterogeneous mixture properties by arranging the data in groups having the same patterns or rules. However, since there are a huge number of possibilities (sometimes infinite) for the data grouping options, it is in reality impossible to verify each and every candidate. The following three issues are of importance in arranging the data into several groups.

1) Number of groups (How much the data is mixed)
2) Method of grouping (How the data is grouped)
3) Appropriate choice of prediction model according to the properties of each group

These issues cannot be solved independently or by following the order from 1) to 3), but they should be solved simultaneously by considering their mutual dependences. For example, when the hypothesis is that data contains a mixture of nonlinear and linear relationships (Fig. 1, Left), a highly accurate prediction model can be obtained by grouping the data into two groups (ellipse B and ellipse A). However, when the hypothesis is that the data contains a mixture of multiple linear relationships (Fig. 1, Right), the optimum number of groups becomes 3. In both left and right parts of Fig. 1, the grouping methods (ellipses) are determined by the sets of data to which the linear (or nonlinear) relationships (prediction models) are applicable, and this fact means that it is not possible to determine 2) by ignoring 1) and 3).

It is obligatory then to consider issues 1) to 3) simultaneously, which is the specific number of data grouping candidates. As an example, let us assume a case in which big data storage of a large volume of sensor and electricity demand data is analyzed to detect the hidden rules. Furthermore, to clarify the essence of this issue, we will limit the candidates for the prediction model (electricity demand prediction formula) to those that can be expressed by a quadratic expression of the explanatory variables (sensor values). When the number of explanatory variables (number of sensors) is fixed at 10, the number of sensors usable in the prediction model at 3 and the number of groups obtained by data grouping at 4, the number of prediction model candidates is calculated approximately at $({}_{100}C_{3})^{4} = 6.84\times10^{20}$ (10^{20} is equal to 1 trillion multiplied by 100 millions). In more complicated cases, there are almost infinite combination candidates of data groups and prediction models. This means that the time taken for a search is at an unrealistic level if simple algorithms are used.

As described in section 1, the solution most often adopted hitherto to solve such a problem was to define the factors altering the rules via trial and error based on expert knowledge and to classify the data accordingly in order to enable the automatic extraction of a single rule for each group. However, to determine the optimum data grouping method for data acquired from such a complex system is very difficult to achieve, even for experts. Constraints are posed by a reduction in the prediction accuracy due to inappropriate grouping and by the huge amount of labor required for the trial and error procedures needed to find the optimum grouping method.

3. Data Mining Based on Heterogeneous Mixture Learning

NEC has developed a new heterogeneous mixture learning technology for use in mining heterogeneous mixture data. This technology is capable of the high speed optimization of the three issues 1) to 3) referred to in section 2 above by avoiding issues related to data grouping or a sudden increase in prediction model combinations.

Below, we explain the differences between learning with the previous techniques (such as the cross-validation or the Bayesian information criterion) and the heterogeneous mixture learning as shown in **Fig. 2** . Previous techniques calculated the scores (information criteria) for the model candidates and selected the model with the best score. However, as we described in section 2 above, an unrealistic calculation time would be required if these techniques were applied to the learning of heterogeneous mixture data due to the enormous number of model candidates. On the other hand, heterogeneous mixture learning is capable of adaptive searching of issues 1) to 3), which are the number of groups, the method of grouping and the prediction model for each group. This makes it possible to find the optimum data grouping and prediction model by investigating models with high prediction accuracies without searching unpromising candidates.

The advanced search and optimization of the heterogeneous mixture learning is backed by the latest machine learning theory called "factorized asymptotic Bayesian inference" [2)3)4)] .

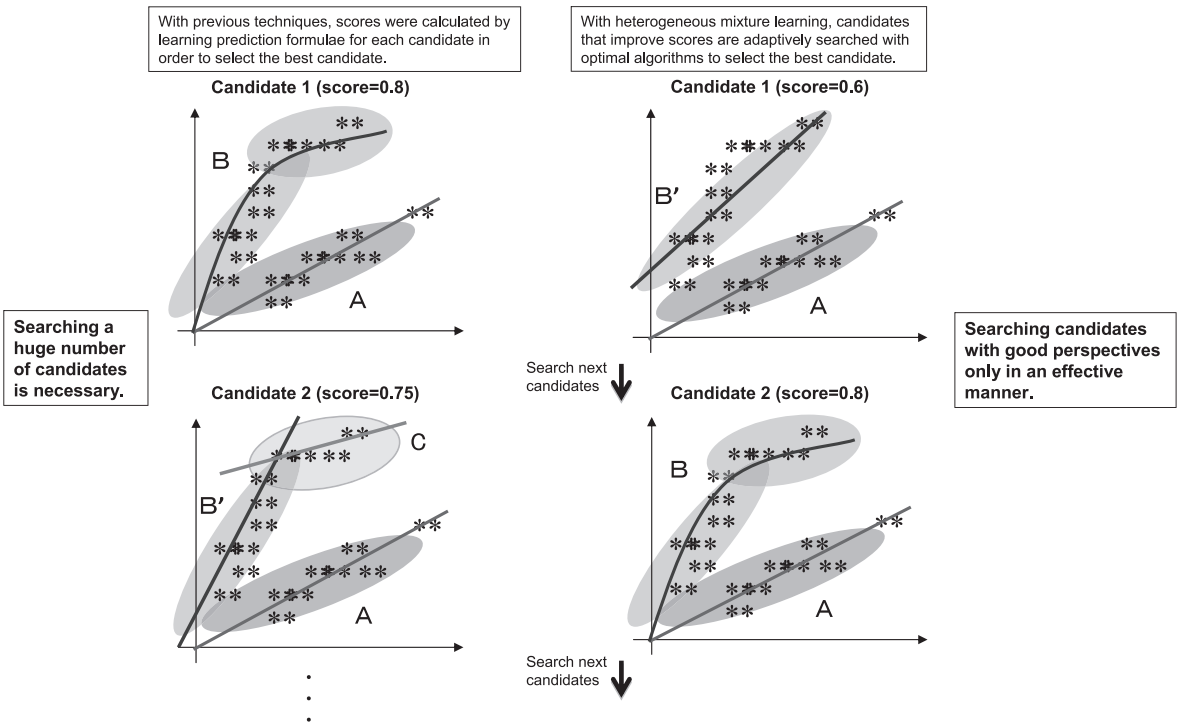


Fig. 2 Differences in data grouping and prediction model search methods between heterogeneous mixture learning and previous techniques.

The foundation of the heterogeneous mixture learning is supported by the following properties of this theory.

(1) Factorized information criterion:
In the field of machine learning models that have multiple modes are called "non-regular (singular) models." It is known that with these models it is impossible to group data optimally or measure the advantages of prediction models using the traditional information criteria (such as the Bayesian information criterion or Akaike's information criteria). In contrast, the heterogeneous mixture learning can group data optimally and select the appropriate prediction model by applying an original criterion, called the factorized information criterion, to non-regular models.

(2) Adaptive search algorithms:
As explained in Fig. 2, the heterogeneous mixture learning searches data and optimally varies the number of groups, the grouping method and the prediction model for each group. It uses a special technique in the search to ensure that after change a model is always superior to the previous model in terms of the factorized information criterion. The possibility of adaptive selection of a model that is always superior to the previous model means that there is no need to search models that are inferior to the previous model and that data grouping and prediction model finding from the large amount of candidates is quicker.

(3) Elimination of adjustment parameters (dependencies on individual skills):
Many of the machine learning and data mining algorithms contain parameters that should be adjusted manually by the analyzer. Nevertheless, this adjustment requires mathematical understanding of the algorithms and consequently very advanced skills in general. Meanwhile, the heterogeneous mixture learning determines the previously required adjustment parameters by means of the factorized asymptotic Bayesian inference. This makes it possible to automate the analysis by eliminating dependencies on individual skills.

(4) Identification of models:
As shown in the left and right parts of Fig. 1, the data grouping and prediction model candidates contain models with very close (or completely equivalent) performances, and it is known that the presence of equivalent models causes issues in learning the models (the problem of non-identification of models) [5].
It has theoretically been indicated that the heterogeneous mixture learning has a "model identification performance" that can identify models uniquely under circumstances in which equivalent models are present.

4. Demonstration Experiment on Electricity Demand Prediction

To confirm the effects of a heterogeneous mixture model, we conducted a demonstration experiment on the prediction of electricity demand of a building. Backed by the recent worldwide rise in fuel prices, accurate electricity demand predictions and the application of more intelligent controls beyond simple peak cutting are expected to contribute to important energy cost reductions.

Fig. 3 shows the prediction by heterogeneous mixture learning (top row), simple prediction (prediction with the traditional machine learning technology, middle row) and the prediction by manual data grouping (prediction assuming that the electricity demand trend varies depending on the day of week and a model is learned for every day of the week, bottom row). Observation of the prediction errors (deviation of lines) of these techniques shows that the sections with large prediction errors (enclosed in circles or ellipses) exist in the middle and bottom rows due to incorrect learning of the points where the demand trend switches. On the other hand, with the heterogeneous mixture learning, several models are switched automatically (in this case, the data is arranged into three groups), indicating that accurate predictions are possible even at the points where the trend switches. The heterogeneous mixture learning succeeded in improving the prediction accuracy by 7.6 points (10.3% → 2.7%) compared to the simple prediction by 2.1 points (4.8% → 2.7%) compared to the manual mode.

5. Broad Applicability

When these technologies are applied for example to the prediction of electricity demand of a building, highly accurate prediction becomes possible by extracting and utilizing various rules mixed in the collected data, even if the relationships between electricity demand and factors such as surrounding temperature, day of the week or time of day of the building are constantly changing. These technologies may also be useful in the medical field for detecting abnormal patterns from "lifelog" data, potentially resulting in the early detection of asymptomatic illnesses.

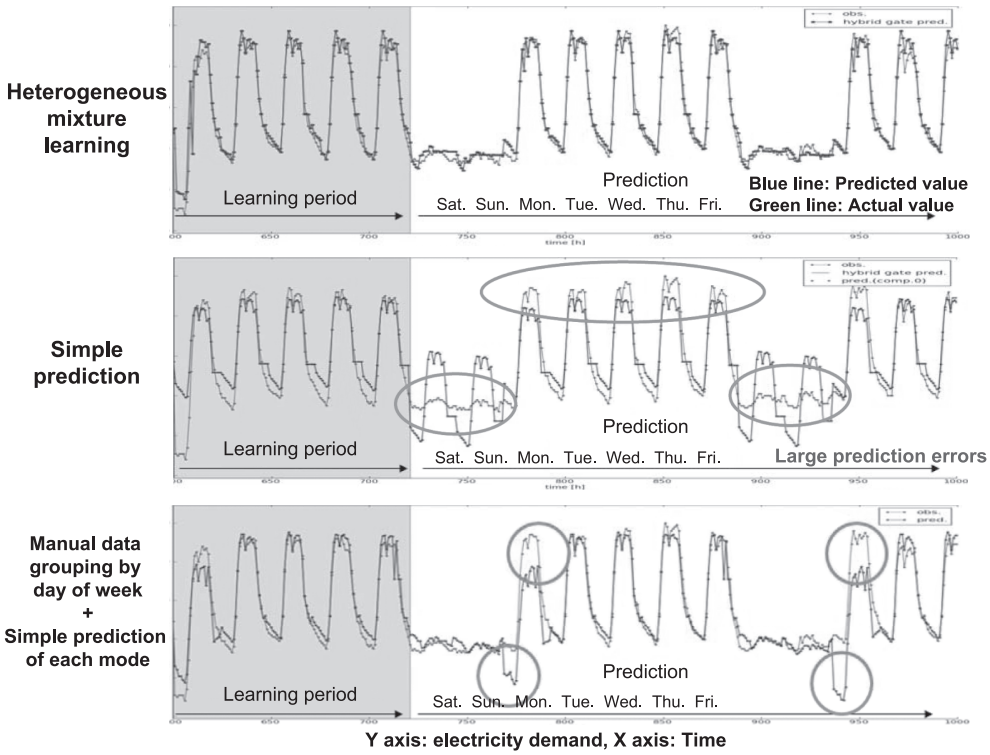


Fig. 3 Results of the application of heterogeneous mixture learning on building electricity demand prediction.

6. Conclusion

The heterogeneous mixture learning technology is an advanced technology used in big data analysis. In the above, we introduced difficulties that are inherent in heterogeneous mixture data analysis, the basic concept of heterogeneous mixture learning and the results of a demonstration experiment that dealt with electricity demand predictions.

As the big data analysis increases its importance, heterogeneous mixture data mining technology is also expected to play a significant role in the market. The range of application of heterogeneous mixture learning will be expanded broader than ever in the future.

References

1) Yano Research Institute: “2012 Big Data Market - Possibilities and strategies of newcomers -,” C54101300, 2012.4
2) R. Fujimaki, S. Morinaga: “Factorized Asymptotic Bayesian Inference for Mixture Modeling,” JMLR W&CP 22: 400-408, 2012
3) R. Fujimaki, K. Hayashi: “Factorized Asymptotic Hidden Markov Models,” Proceedings of the 29th International Conference on Machine Learning (ICML), 2012
4) R. Fujimaki, Y. Sogawa, S. Morinaga: “Online heterogeneous mixture modeling with marginal and copula selection,” Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.645-653, 2011
5) S. Watanabe: “Algebraic geometry and statistical learning,” Cambridge University Press, 2009

The Most Advanced Data Mining of the Big Data Era

Authors' Profiles

FUJIMAKI Ryohei
Doctor of Engineering
Research Member
Media Analytics
NEC Laboratories America

MORINAGA Satoshi
Doctor of Engineering
Principal Researcher
Knowledge Discovery Research Laboratories
Central Research Laboratories

The details about this paper can be seen at the following.

Related URL:

NEC R&D on Data Mining:
http://www.nec.co.jp/rd/en/datamining/index.html

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.7 No.2 Big Data

Vol.7 No.2
September, 2012

Special Issue TOP