Formalizing expert knowledge through machine learning

Tsuyoshi Idé

Abstract This chapter addresses one of the key questions in service science: how to formalize expert knowledge. While this question has been treated mainly as a task of formal language design, we use an alternative approach based on machine learning. Investigating the history of expert systems in artificial intelligence, we suggest that three criteria, generalizability, learnability, and actionability, are critical for extracted expert rules. We then conclude that machine learning is a promising tool to satisfy these criteria. As a real example, we perform a case study on a task of condition-based maintenance in the railway industry. We demonstrate that our proposed statistical outlier detection method achieves good performance for early anomaly detection in wheel axles, and thus in encoding expert knowledge.

Key words: knowledge acquisition bottlenecks, machine learning, sensor data, anomaly detection, dependency discovery

1 Introduction

Service industrialization in traditional non-service industries is a recent major trend in the global economy. Related to this, one of the recent accomplishments in service science research is the establishment of the concept of value co-creation between different entities in service systems as a universal view that applies even to newly service-industrialized industries [12]. It is therefore interesting to study in what sense the traditional business domains that have been thought of as non-service industries can be understood in terms of value co-creation.

This chapter focuses on the Japanese railway industry, which has some of the highest service standards in the world for such metrics as on-time operations [15].

Tsuyoshi Idé
IBM Research – Tokyo, 5-6-52 Toyosu, Kōtō-ku, 135-8511 Tokyo, Japan, e-mail: goodidea@jp.ibm.com

While the high service quality can be thought of as the result of a value co-creation process involving expert engineers, little is known about the mechanisms of the process from service science perspectives. Thus studying the value co-creation processes of the field engineers is of particular interest.

In general, expert engineers make decisions based on their experience and observations. For example, an engineer may tap a bogie with a hammer, and carefully listen to the sound to see if the system is working properly. Another engineer may check multiple different sensor values to decide on an action for a malfunction. Although much of the maintenance work is well documented, formalizing the subtle decision-making processes in such situations is generally very hard. Perhaps the difficulty of documentation itself is the source of differentiation, and, if this is the case, then transforming such expert knowledge into formalized knowledge amounts to analyzing the value creation process itself.

Therefore, we consider the problem of *how to formalize expert knowledge* as a particularly important problem in service science. Taking condition-based maintenance (CbM) in the railway industry as an example, we give a case study to suggest one possible solution to the problem.

1.1 Condition-based maintenance in the railway industry

In the railway industry, the basic strategy of maintenance is preventive maintenance, where some action is taken before an accident occurs. In modern preventive maintenance, time-based maintenance is still the mainstream approach, where periodic replacements of parts are done based on the predefined "safe" lifetimes of the individual parts. A general trend is that a shift from time-based maintenance is taking place, moving to condition-based maintenance, where the individual parts are only replaced based on their actual conditions as measured by sensors.

Although the advantages of CbM are clear in terms of cost savings and safety, CbM requires sophisticated analytics technologies to assess the health of the system from the sensor data. For example, in a shinkansen car, each of the journal boxes (wheel axle boxes) is equipped with a thermal senor, and the train is designed to make an emergency stop if the temperature exceeds 140 °C [10]. However, it is known that the temperature data is heavily influenced by external conditions such as outdoor weather, and that anomalies manifest themselves in many ways. As a result, early anomaly detection for CbM is extremely difficult unless the anomaly is simple, as in the example of the 140°C threshold. To consider a real example, it has been extremely difficult for existing technologies to distinguish between temperature decreases due to rain and temporal temperature decreases by oil leakage due to lubrication failure. Therefore, post-run maintenance checks by experienced engineers is almost the only option at this time.

1.2 Goal of this chapter

Our task is to study how to extract such expert knowledge in the form of reusable rules. For this goal, we have two basic problems:

- What kind of language is appropriate for knowledge representation?
- How can we construct useful rules from experience?

As discussed later, for the first problem, we argue that the traditional assumption that natural language is always valid for knowledge representation is not necessarily true. The implication of this can be profound in service science, since it might entail a paradigm shift just like the one when mathematical astronomy drove out metaphysical studies based on ancient stories and myths.

For the second problem, we argue that machine learning, which is essentially data-driven, is the most appropriate approach. Since a service system involves a value co-creation process by different entities through various interactions, the system is almost always complex. For complex systems, relying upon knowledge on the microscopic models of the system is unrealistic. In this sense, data-driven approaches including machine learning and data mining are of particular importance in service science.

The layout of this chapter is as follows. In the next section, we briefly take a look at the history of expert systems in artificial intelligence. In Section 3, we go through the basic strategy of machine learning to see the importance in the service science research. In Section 4, we introduce a machine learning approach to encode the expert knowledge within a probabilistic model. In Section 5, we present a detailed case study from the railway industry. Finally, Section 6 summarizes this chapter.

2 A brief history of artificial intelligence: failures and successes of expert systems

This section briefly reviews the history of expert systems in artificial intelligence (AI). An expert system is a database (DB) system composed of a knowledge base and a search engine that traverses the knowledge base to identify the most appropriate answers to queries. The first expert systems appeared in the late 1970s. While expert systems have never been extensively used in the real world, a recent success, the victory of a DeepQA system in an American quiz show, gives us useful insights into our problem of how to formalize expert knowledge.

2.1 The failure: the knowledge acquisition bottleneck

The original approach in computer science to formalizing knowledge was to express the knowledge in a formal language such as Prolog, and to accumulate the rules in

a DB. Expert systems were made of such DBs [13], and IF-THEN rules can be viewed as the simplest example of the formal language. In MYCIN [3], which is undoubtedly the best-known expert system, a typical rule looks like this:

```
IF the identity of the germ is not known with certainty
AND the germ is gram-positive
AND the morphology of the organism is "rod"
AND the germ is aerobic
THEN there is a strong probability (0.8) that the germ is of type enterobacteriacae
```

Since traversing the rule DB is often time-consuming, a major research focus was put on search and enumeration technologies for the rules.

Whatever formal language is used in an expert system, the assumption is that natural language is always a valid representation of human knowledge. Since our thoughts are tightly connected with our languages, using natural languages and their variants such as Prolog has been thought of as a literally natural approach.

Despite the serious and extensive research invested in it, MYCIN has never actually been used in practice [14]. There are at least two reasons. First, the limitation of computational resources was an issue in performing DB search on a realistic time scale and data volume. Second, and perhaps most importantly, MYCIN could not produce meaningful answers unless seemingly complete knowledge was available in advance. The technical highlight of MYCIN was its algorithm for computing the value of the confidence of the rules. While it worked well when a rich knowledge base was available, it was not very useful in most real-world cases, where only an incomplete set of knowledge is available.

If knowledge acquisition is a problem, then how can we acquire it at a minimum cost? In spite of extensive effort in the AI community for decades, no conclusive answer was ever obtained at least not in the way originally imagined. This is the well-known problem of the *knowledge acquisition bottleneck*.

2.2 The success: the victory of DeepQA

In 2011, we witnessed IBM's "DeepQA" system [6] beating human champions in playing an American TV quiz show. This was really epoch-making news in the history of AI, and perhaps in the history of service science. Question-answer (QA) systems are one type of expert systems, where each query is processed to list the candidate answers, and such systems have been one of the major research topics in AI. However, the task DeepQA addressed was slightly but significantly different from the traditional problem setting in that the DeepQA system is capable of handling *open-domain* questions.

In the traditional problem setting, a QA system is assumed to handle queries within a closed domain. In a sense, traditional QA systems are straightforward machines, which tell us only the anticipated answers. However, in quiz shows, the variety of answers is almost infinite. Also, the system must handle queries that are

far from formal language. A query may contain puns and metaphors, and understanding the query itself is challenging. In this sense, an open-domain QA system is an expert system that is capable of searching over an infinite space. This is why the victory of DeepQA is so epochal.

We note that the DeepQA system does not rely on an integrated ontology in the DB. Such an approach was not appropriate for the open-domain QA task. In fact, in the DeepQA system, individual rules are shallow and partial, and statistical machine learning integrates them into a single QA system [6].

2.3 Implications to service science

What is the implication of the failure and success of expert systems to service science? In the Introduction, we discussed that formalizing expert knowledge amounts to analyzing the value co-creation process in recently service-industrialized domains. As is understood from the example of CbM in the railway industry, an expert must handle open-domain questions, and must be capable of updating his/her knowledge based on newly observed facts. To summarize our claim:

Claim 1 *In service science, formalizing expert knowledge is one of the key problems to understand the value co-creation process. Our goal is to capture the rules of decision patterns of experts so that three criteria are satisfied:*

- *Generalizability*
- *Learnablity*
- *Actionability*

For generalizability, the rule must handle unseen situations by generalizing a finite amount of previously observed data. This might look like a leap in logic since we need to handle infinite situations based on a finite data set. However, as in DeepQA, statistical machine learning allows us to generalize the knowledge through statistical abstraction.

For learnablity, in order to address the knowledge acquisition bottleneck, we need functions of capturing previous experiences into the system, and updating the current decision rule with newly acquired knowledge if needed.

For actionability, the captured rule must not be a black box, and the decision rule must provide understandable information to humans. In the case of DeepQA, a list of candidate answers is given together with confidence values. Also, in machine learning, a decision rule is always given as a (possibly nonlinear) function of different features. By looking at the weight of each feature, one can get some insight into the system for which factors play critical roles.

3 Knowledge acquisition in service systems: strategy of machine learning

This section points out the importance of machine learning in service science. We first argue that natural language is not necessarily the only choice for knowledge representation (Claim 2). Then we point out that the data-driven approach in machine learning is of critical importance for modeling service systems, where complex interactions between entities are involved (Claim 3).

3.1 Functional relationship as generalized rule

As discussed in Subsection 2.1, the failure of traditional expert systems posed a challenge to the validity of natural language as a format for knowledge representation. As a concrete example, we focus on a task of CbM in the railway industry. Our goal is early anomaly detection for wheel axles by using temperature sensors. For example, in the shinkansen trains, there are sixteen cars in each set, and each car has eight journal boxes (on both sides of the four wheel axles). Our goal is to detect early signs of anomalies by analyzing the 128-dimension temperature data.

In predictive maintenance, the boundary between normal and abnormal conditions is not clear. Thus it is fair to define our problem as computing the anomaly score for each wheel axle. Let y_i be the anomaly score for the i-th axle box. If our output is y_i for all is, then our input variable is the temperature data. We express one measurement of the temperature sensors as a 128-dimension (column) vector

$$\boldsymbol{x} \equiv [x_1, \ldots, x_i, \ldots, x_M]^\top,$$

where x_i denotes the temperature of the i-th axle box, and M is a generalized notation representing the total number of wheel axle boxes (in the Shinkansen case, $M = 128$). Now our problem is to compute y_i ($i = 1, 2, \ldots, M$) for a given $\boldsymbol{x}$, based on the previous measurements $\mathscr{D}$ under normal conditions.

If we employ an IF-THEN encoding, the decision rule may look like

`IF` x_{32} is 10 °C greater than any other axle box
`THEN` $y_{32} = 1$
`ELSE` $y_{32} = 0$.

However, this type of expression does not work well in practice. The data is noisy, and includes a lot of outliers mainly due to the effects of external factors such as the weather. The degree of the external effect may be different depending on the location of the wheel axles. For example, the influence of the wind is more critical for the first and last cars in high-speed trains, and the temperatures in those cars tend to be much lower than in the other cars. In this way, there is an almost infinite number of factors that may affect the rules, and handling such exceptions with fixed IF-THEN rules is effectively impossible.

If we state the problem in the most general manner, our goal is to obtain the function f_i such that

$$y_i = f_i(\boldsymbol{x}|\mathscr{D})$$

for $i = 1, 2, \ldots, m$, or

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}|\mathscr{D})$$

in the vector form combining all of the M relationships. Here '|' represents 'conditioned on' or 'given.' In this case, the function $f_i(\boldsymbol{x}|\mathscr{D})$ encodes expert knowledge, which can be thought of as a generalization of natural-language-based rules. Starting from this general expression, we determine an optimal functional relationship based on the data. We call this type of function a rule, assuming a mathematically well-defined functional relationship. In other words, we put more weight on mathematics as a language to describe service systems. This strategy reminds us of Galileo's famous statement [5]

> Philosophy is written in this grand book, the universe.... It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures;....

Galileo was one of the first modern thinkers to clearly state that the laws of nature are mathematical. We believe that Galileo's statement is at least partly true in service systems. We summarize our strategy towards expert knowledge formalization.

Claim 2 *A rule is a functional relationship between a decision variable* $\boldsymbol{y}$ *and observables* $\boldsymbol{x}$*. The function, which is assumed to be a mathematical relationship in general, is to be optimally determined from the data* $\mathscr{D}$ *so as to best satisfy the criteria in Claim 1.*

3.2 Data-driven approach to rule induction

Now our problem is how to determine the function $f_i(\boldsymbol{x}|\mathscr{D})$ from the data. For this purpose, we employ a machine-learning approach. Machine learning and data mining are relatively new academic disciplines originally intended to address the knowledge acquisition bottleneck [11].

To be concrete, again, consider the task of CbM for wheel axles. Our goal is to compute the anomaly score y_i for each axle, based on the data $\mathscr{D}$ containing N samples in the past under normal conditions. To represent explicitly, $\mathscr{D}$ is written as

$$\mathscr{D} \equiv \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(N)}\}. \tag{1}$$

In the modern theory of statistical machine learning, this function is determined based on probabilistic distributions. In a typical formulation, we build a probability distribution of $\boldsymbol{x}$ from $\mathscr{D}$. Let $p(\boldsymbol{x}|\mathscr{D})$ be such a distribution. Since this function represents the likelihood of an observation $\boldsymbol{x}$, we see that

- If $p(\boldsymbol{x}|\mathscr{D})$ is large, then the sample $\boldsymbol{x}$ takes a value close to its expected value. The state of the system is expected to be in a normal condition, and the anomaly score should be small.
- If $p(\boldsymbol{x}|\mathscr{D})$ is close to zero, then the sample $\boldsymbol{x}$ takes a value far from its expected value. The state of the system is expected to be in an anomalous condition, and the anomaly score should be large.

This assumes a uni-modal distribution. We will give an explicit mathematical expression for the anomaly score later.

Note that this type of approach focused only on the functional relationship between the input and output. The system is treated as a black box, and no attention is paid to precisely modeling the microscopic mechanism of the system. For example, in a later section, we give a quadratic form of function for the anomaly score. However, we do not mean the physical mechanism of the system is represented as a quadratic function of the temperatures. Unless we have complete knowledge of the system, which is unlikely in many real-world cases, we believe that the data-driven approach of machine learning is quite reasonable.

While this statement is primarily for physical systems, the situation is parallel in service systems. Service science is an academic discipline that studies the process of value co-creation between different entities [12], and one of the recent research focuses is on holistic service systems containing interacting entities and value exchange mechanisms. Since holistic service systems are complex interacting systems, analytic approaches involving reduction to elements are not always effective in studying the processes of value creation. If we think of the value created as an output of a holistic service system, the a data-driven approach looks promising. Based on pervious observations, machine learning would allow us to build a predictive model of the output, and to clarify what kind of factors play important roles in the value creation.

To summarize, our point is that:

Claim 3 *Service systems involving value co-creation process between different entities is complex interacting systems in nature. In modeling the value co-creation process, data-driven approach is promising. In this sense, machine learning is one of the most important disciplines in service science.*

4 Dependency-based statistical anomaly detection

This section describes a statistical anomaly detection method to encode experts' logic for anomaly detection. Our task is early fault detection from the temperature sensors of journal boxes. As discussed earlier, the measurements are affected by a lot of external disturbances such as the effect of the weather, and using a fixed threshold for individual axles is not an optimal approach. Also, since individual train cars are not identical to each other, the behavior of individual journal box temperatures can differ. Our main technical challenge is minimize these sources of

confusion. In this section, we will show that a subspace extraction technique and sparse structure learning solve these hard problems.

In what follows, we denote our input (wheel axle temperature) by an M-dimensional vector $\boldsymbol{x} \in \mathbb{R}^M$. We assume that we are given a data set as in Eq. (1), and that $\mathscr{D}$ has been standardized to have zero mean and unit variance.

4.1 Subspace extraction technique

Let us focus on how to suppress the unwanted effects of the weather. Since our metric is temperature, we expect that the value returned by a sensor gets smaller when it is rainy or windy, and increases when it is sunny. This means that suppressing this effect amounts to extracting the primary trend of the data. To find the primary trends of the temperature vectors $\{\boldsymbol{x}^{(n)}\}$, we consider an optimization problem:

$$\mathsf{W} = \arg\max_{\mathsf{W}} \sum_{n=1}^{N} \sum_{i=1}^{d} (\boldsymbol{w}_i^\top \boldsymbol{x}^{(n)})^2 \quad \text{subject to } \boldsymbol{w}_i{}^\top \boldsymbol{w}_j = \delta_{i,j}, \tag{2}$$

where we denote the primary direction as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d$, and

$$\mathsf{W} \equiv [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d] \ \in \ \mathbb{R}^{M \times d}.$$

Using Lagrange's coefficient α_i for the i-th constraint, we straightforwardly obtain the equation

$$\mathsf{S}^0 \boldsymbol{w}_i = \alpha_i \boldsymbol{w}_i, \ \ i = 1, \cdots, d,$$

where the (i, j)-element of the sample covariance matrix $\mathsf{S}^0 \in \mathbb{R}^{M \times M}$ is given by

$$\mathsf{S}^0_{i,j} \equiv \frac{1}{N} \sum_{n=1}^{N} x_i^{(n)} x_j^{(n)}, \tag{3}$$

which is the same as the correlation coefficient matrix for this data. These equations state that the directions $\boldsymbol{w}_i$ are the eigenvectors of S^0.

We now re-normalize the original data $\boldsymbol{x}$ by subtracting the primary directions as

$$\boldsymbol{\xi} \equiv (\mathsf{I} - \mathsf{W}\mathsf{W}^\top)\boldsymbol{x}. \tag{4}$$

The number of the eigenvectors, d, is a parameter determined by trial and error.

4.2 Graphical Gaussian models

Next we consider how to analyze the dependencies between variables. Figure 1 illustrates the general approach. Given the data $\mathscr{D}$, our goal is to find a graph rep-

resenting any hidden dependencies among the variables. Since we are interested in a major structure that would not be affected by the noise, it is important to obtain a *sparse* graph.

To model the dependency graph, we use a graphical Gaussian model (GGM). In our problem, where the temperatures of the journal boxes are monitored, the physical behaviors of the variables are expected to be similar, and thus the variables are expected to be highly correlated with each other. Since the correlation between the variables is a natural statistic of Gaussian, a multivariate Gaussian distribution is a reasonable choice for modeling the system. The GCM is the simplest graphical model based on multivariate Gaussian.

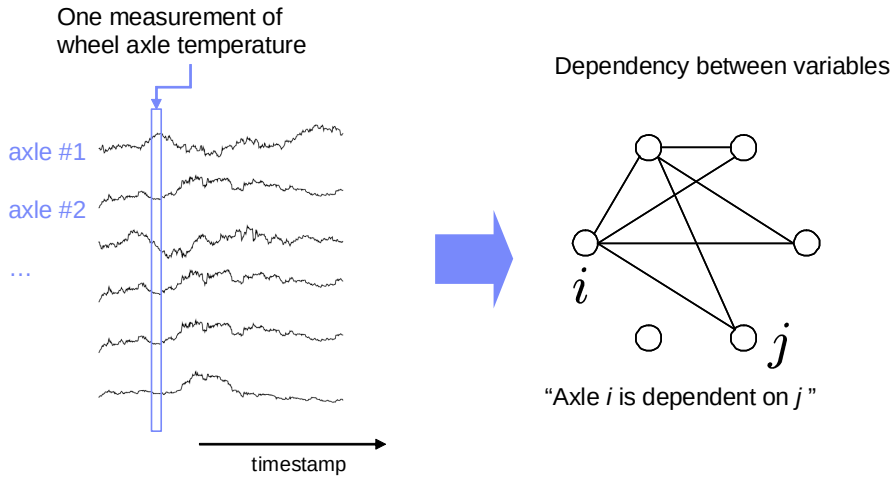

Fig. 1 Sparse structure learning finds sparse dependencies between variables in the data.

For a zero-mean M-dimensional random variable $\boldsymbol{\xi} \in \mathbb{R}^M$, the GGM assumes an M-dimensional Gaussian distribution

$$\mathcal{N}(\boldsymbol{\xi}|\mathbf{0},\Lambda^{-1}) = \frac{\det(\Lambda)^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^\top \Lambda \boldsymbol{\xi}\right), \tag{5}$$

where det represents the matrix determinant, and $\Lambda \in \mathbb{R}^{M\times M}$ denotes a precision matrix. We denote by $\mathcal{N}(\cdot|\boldsymbol{\mu},\Sigma)$ a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. The precision matrix is formally defined as the inverse of a covariance matrix.

In the GGM, a Gaussian distribution is associated with a graph (V,E), where V is the set of nodes containing all of the M variables, and E is a set of edges. The edge between ξ_i and ξ_j is absent if and only if they are independent conditioned on all of the other variables. Under the Gaussian assumption, this condition is represented as

$$\Lambda_{i,j} = 0 \quad \Rightarrow \quad \xi_i \perp\!\!\!\perp \xi_j \mid \text{ other variables}, \tag{6}$$

where $\perp\!\!\!\perp$ denotes statistical independence.

The condition (6) can be most easily understood by explicitly writing down the conditional distribution. Let us denote $(\xi_i, \xi_j)^\top$ as $\boldsymbol{\xi}_a$, and the rest of the variables by $\boldsymbol{\xi}_b$. For centered data, a standard partitioning formula of Gaussian (see, e.g. [2], Sec. 2.3) gives the conditional distribution as

$$p(\boldsymbol{\xi}_a|\boldsymbol{\xi}_b) = \mathcal{N}(\boldsymbol{\xi}_a| - \Lambda_{aa}^{-1}\Lambda_{ab}\boldsymbol{\xi}_b, \Lambda_{aa}^{-1}), \tag{7}$$

where, corresponding to the partitioning between $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_b$, we put

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}. \tag{8}$$

In this case, Λ_{aa} is 2×2, so the inverse can be analytically calculated, giving the off-diagonal element proportional to $\Lambda_{i,j}$. Thus if $\Lambda_{i,j} = 0$, then x_i and x_j are statistically independent conditioned on the rest of the variables.

Our goal in this subsection is to find a sparse Λ whose entries are nonzero for essentially coupled pairs and zero for weakly correlated pairs that might be spuriously created by noise. Such a sparse Λ will represent an essential dependency structure not due to noise, and thus should be useful for detecting correlation anomalies. In real-world noisy data, however, every entry in the transformed sample covariance matrix

$$\mathsf{S}_{i,j} \equiv \frac{1}{N}\sum_{n=1}^{N} \xi_i^{(n)} \xi_j^{(n)} \tag{9}$$

will be nonzero, and the precision matrix Λ will not in general be sparse. Here

$$\boldsymbol{\xi}^{(n)} \equiv (\mathsf{I} - \mathsf{W}\mathsf{W}^\top)\boldsymbol{x}^{(n)}.$$

Moreover, if there are highly correlated variables, S will tend to become rank deficient, and Λ will not even exist. Even if S is full rank in theory, it is sometimes the case that matrix inversion is numerically unstable when M is more than several tens. This is an essential difficulty in traditional covariance selection procedures [4], where small entries in Λ are set to be zero starting from the smallest. Since our assumption is that the data include some highly correlated variables, which holds very generally for sensor data, such approaches are of little use in our context. This motivates us to use an L_1-penalized maximum-likelihood approach.

4.3 Sparse structure learning

In the GGM, structure learning is reduced to finding a precision matrix Λ for the multivariate Gaussian. If we ignore any regularization, we can get Λ by maximizing the log-likelihood

$$\ln \prod_{t=1}^{N} \mathcal{N}(\boldsymbol{\xi}^{(t)} | \mathbf{0}, \Lambda^{-1}) = \text{const.} + \frac{N}{2} \left\{ \ln\det(\Lambda) - \text{tr}(\mathsf{S}\Lambda) \right\},$$

where tr represents the matrix trace (the sum over the diagonal elements). If we use the well-known formulas for matrix derivatives

$$\frac{\partial}{\partial \Lambda} \ln\det(\Lambda) = \Lambda^{-1}, \quad \frac{\partial}{\partial \Lambda} \text{tr}(\mathsf{S}\Lambda) = \mathsf{S}, \tag{10}$$

then we readily obtain the formal solution $\Lambda = \mathsf{S}^{-1}$. However, as mentioned before, this produces a smaller amount of practical information on the structure of the system, since the sample covariance matrix is often rank deficient and the resulting precision matrix will not in general be sparse.

Therefore, instead of the standard maximum likelihood estimation, we solve an L_1-regularized version of the maximum likelihood:

$$\Lambda^* = \arg\max_{\Lambda} f(\Lambda; \mathsf{S}, \rho), \tag{11}$$

$$f(\Lambda; \mathsf{S}, \rho) \equiv \ln\det\Lambda - \text{tr}(\mathsf{S}\Lambda) - \rho ||\Lambda||_1, \tag{12}$$

where $||\Lambda||_1$ is defined as $\sum_{i,j=1}^{M} |\Lambda_{i,j}|$. Thanks to the penalty term, many of the entries in Λ will be exactly zero. The penalty weight ρ is an input parameter, which works as a threshold below which the correlation coefficients are thought of as zero.

Since Eq. (11) is a convex optimization problem [1], we can use subgradient methods to solve it. Recently, Friedman, Hastie, and Tibshirani [8] proposed an efficient subgradient algorithm named graphical lasso. We describe briefly in this subsection.

The graphical lasso algorithm first reduces the problem Eq. (11) to a series of related L_1-regularized regression problems by utilizing a block coordinate descent technique [1, 7]. Using Eq. (10), we see that the gradient of Eq. (11) is given by

$$\frac{\partial f}{\partial \Lambda} = \Lambda^{-1} - \mathsf{S} - \rho \ \text{sign}(\Lambda), \tag{13}$$

where the sign function is defined so that the (i, j) element of the matrix $\text{sign}(\Lambda)$ is given by $\text{sign}(\Lambda_{i,j})$ for $\Lambda_{i,j} \neq 0$, and a value $\in [-1, 1]$ for $\Lambda_{i,j} = 0$.

To use a block coordinate descent algorithm to solve $\partial f / \partial \Lambda = 0$, we focus on a particular single variable x_i, and partition Λ and its inverse as

$$\Lambda = \begin{pmatrix} L & \boldsymbol{l} \\ \boldsymbol{l}^\top & \lambda \end{pmatrix}, \quad \Sigma \equiv \Lambda^{-1} = \begin{pmatrix} W & \boldsymbol{w} \\ \boldsymbol{w}^\top & \sigma \end{pmatrix}, \tag{14}$$

where we assume that the rows and columns are always arranged so that the x_i-related entries are located in the last row and column. In these expressions, $W, L \in \mathbb{R}^{(M-1)\times(M-1)}$, $\lambda, \sigma \in \mathbb{R}$, and $\boldsymbol{w}, \boldsymbol{l} \in \mathbb{R}^{M-1}$. Corresponding to this x_i-based partition, we also partition the sample covariance matrix S in the same way, and write it as

$$\mathsf{S} = \begin{pmatrix} S^{\setminus i} & \boldsymbol{s} \\ \boldsymbol{s}^\top & s_{i,i} \end{pmatrix}. \tag{15}$$

Now let us find the solution of the equation $\partial f/\partial \Lambda = 0$. Since Λ must be positive definite, the diagonal elements must be strictly positive. Thus, for the diagonal elements, the condition of the vanishing gradient leads to

$$\sigma = s_{i,i} + \rho. \tag{16}$$

For the off-diagonal entries represented by $\boldsymbol{w}$ and $\boldsymbol{l}$, the optimal solution under which all the other variables are held constant is obtained by solving

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||W^{\frac{1}{2}}\boldsymbol{\beta} - \boldsymbol{b}||^2 + \rho \, ||\boldsymbol{\beta}||_1 \right\} = 0, \tag{17}$$

where $\boldsymbol{\beta} \equiv W^{-1}\boldsymbol{w}$, $\boldsymbol{b} \equiv W^{-1/2}\boldsymbol{s}$, and $||\boldsymbol{\beta}||_1 \equiv \sum_l |\beta_l|$. For the proof, see [9]. This is an L_1-regularized quadratic programming problem, and again can be solved efficiently with a coordinate-wise subgradient method [8].

Now to obtain the final solution Λ^*, we repeatedly solve Eq. (17) for $x_1, x_2, ..., x_M, x_1, ...$ until convergence. Note that the matrix W is full rank due to Eq. (16). This suggests the algorithm is numerically stable. In fact, as shown later, it gives a stable and reasonable solution even when some of the variables are highly correlated.

Once we get the optimal Λ^*, we have the probabilistic model of the data as

$$p(\boldsymbol{x}|\mathscr{D}) = \mathscr{N}(\boldsymbol{\xi}|\boldsymbol{0}, \Lambda^*)$$
$$\boldsymbol{\xi} = \mathsf{I} - \mathsf{W}\mathsf{W}^\top \boldsymbol{x}$$

The next section will define the anomaly score using this model.

4.4 Anomaly score

Now that a complete probabilistic model $p(\boldsymbol{x}|\mathscr{D})$ has been defined, we can proceed to the next step. Here we define the anomaly score for the i-th variable as

$$y_i(\boldsymbol{x}) \equiv -\ln p(\xi_i|\xi_1, .., \xi_{i-1}, \xi_{i+1}, ..., \xi_M, \mathscr{D}). \tag{18}$$

Note that we have M scores, corresponding to individual variables, for a single observation $\boldsymbol{x}$. The definition tells us the discrepancy between the value of the i-th variable and its expected value given the surrounding variables. Thanks to the sparseness, the surrounding variables should be in the same module or cluster as the i-th variable.

Since the right hand side of Eq. (18) is Gaussian, we can analytically write down the expression. For example, for the first variable, the conditional distribution is

$$p(\xi_1|\xi_2,\ldots,\xi_M) = \mathcal{N}\left(\xi_1 \left| -\frac{1}{\Lambda^*_{1,1}}\sum_{i=2}^{M}\Lambda^*_{1,i}\xi_i,\ \frac{1}{\Lambda^*_{1,1}}\right.\right),$$

and the score is given as

$$y_1 \equiv \frac{1}{2}\ln\frac{2\pi}{\Lambda^*_{1,1}} + \frac{1}{2\Lambda^*_{1,1}}\left(\sum_{i=1}^{M}\Lambda^*_{1,i}\xi_i\right)^2. \tag{19}$$

Collecting the M scores into a single vectorial expression, we get the final result of the outlier scores as

$$\boldsymbol{y} \equiv \boldsymbol{y}_0 + \frac{1}{2}\mathrm{diag}(\Lambda^*\boldsymbol{\xi}\mathsf{D}^{-1}\boldsymbol{\xi}^\top\Lambda^*),$$

where $D \equiv \mathrm{diag}^2(\Lambda^*)$ and

$$(\boldsymbol{y}_0)_i \equiv \frac{1}{2}\ln\frac{2\pi}{\Lambda^*_{i,i}}.$$

5 Case study: hot box detection

This section presents experimental results for the two anomaly detection methods introduced in the previous sections. We used these anomaly detection methods with a real problem in the Japanese railway industry.

5.1 Business background

Japanese high-speed train operators have the world's highest service standards for their records of safety and punctuality. Maintaining such a high service standard of service is increasingly difficult due to the growing shortage of skilled engineers. This motivated us to develop a prototype of an anomaly detection system named the IBM Anomaly Analyzer for Correlational Data (ANACONDA).

The task we addressed is often called hot box detection, where the goal is to detect anomalous behaviors of wheel axles based on recorded temperatures. Under normal operations, the temperature of an axle is expected to be highly correlated with the temperatures of the other axles. Thus the dependency-based outlier detection is useful in this application.

Currently, a fixed threshold (typically 140 °C) [10] for individual axles and gear boxes is used based on temperature sensors installed in each of the axle and gear boxes. These sensors, however, are not capable of detecting subtle indications of anomalies, such as a temperature decrease due to oil leakage, which can be particu-

larly hazardous for high-speed trains. To address the limitations of the fixed-sensor approach, frequent manual inspections are required.

Our objective is to enhance the existing system by using additional measurement data. The supplemental anomaly detection system was designed to detect subtle anomalies as indicated by imbalances of the temperature distribution among the axle and gear boxes.

The ultimate goal of the customer is to reduce the human interventions by skilled engineers. Addressing the shortage of skilled engineers is the main concern of the customer.

5.2 Summary of technical challenges

The standard approach to hot box detection is based on temperature monitoring. Axle temperature data has unique characteristics such as being

- Highly dimensional
- Highly correlated
- Strongly and heterogeneously dependent on external conditions

As regards the dimensionality, there are typically eight journal boxes in a single train car, and more than 100 boxes in a complete train. This means that each measurement is a vector of temperatures with more than 100 dimensions.

The most challenging feature is the strong and heterogeneous dependency on external conditions. Our biggest technical challenges were how to filter out the unwanted effects of external conditions:

- How to eliminate the effect of the weather conditions, such as rain, direction and velocity of wind, light intensity, air temperatures, etc.
- How to handle the temperature differences related to car positions because different cars may have different characteristics for their temperatures.
- How to handle temperature differences in axle positions because, even in the same car, different axles may report different temperatures.

As explained before, we handle the first problem by using the subspace extraction method, while sparse structure learning works for the second and third problems.

5.3 Experimental results

We collected a data set $\mathscr{D}$ of axle temperatures under normal conditions, and constructed the model of the system $p(\boldsymbol{x}|\mathscr{D})$, based on the subspace extraction and sparse structure learning techniques. Figure 2 shows an example of a sparse structure found with our approach, where the thickness of each edge represents the amplitude of the corresponding element of Λ^*. The symbols of 2B, 2D, etc. represent

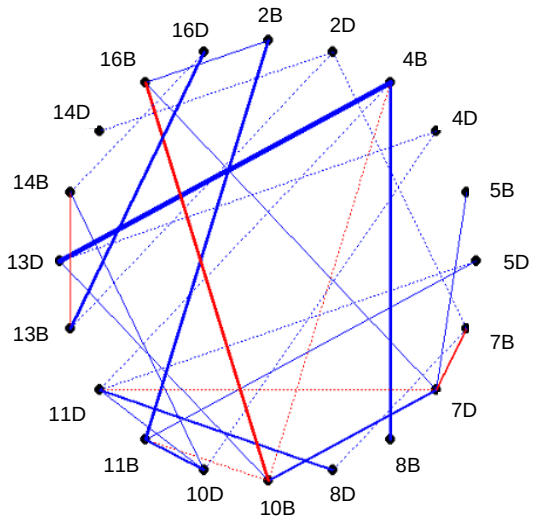

Fig. 2 The dependencies among the variables automatically discovered from the data.

the name of the axle boxes. In spite of the heavy noise in the data, we see that a reasonably sparse graph can be obtained, from which engineers can obtain useful insights into the system. For example, the wheel axle 4B has a strong dependency on 13D. This can be counterintuitive since the car positions are different. However, this result suggests that the fourth and the thirteenth cars have some shared feature, and anomalies can be detected by taking advantage of this feature. This is a good example of how our approach encodes expert knowledge in a concise mathematical expression.

We tested our outlier detection method, and compared the performance with a state-of-the-art method created by domain experts using extensive domain knowledge. The results, shown in Fig. 3, were quite encouraging. We tried two types of preprocessing. In IBM(1), we separated the data set into two portions, each of which corresponds to axle box temperatures on a single side, to create two independent models (i.e. two probabilistic models with $M/2$-dimensional observations). In contrast, in IBM(2) we created a single model.

To evaluate the detection power, we used a separated data set containing anomalous samples. The metric we used is detection power, which is defined for a truly faulty axle i as

$$\frac{1}{\sigma_i}[s_i(\boldsymbol{x}) - \langle y_i \rangle].$$

Here $\langle y_i \rangle$ and σ_i are the mean and the standard deviation of the i-th outlier score over the normal samples in $\mathscr{D}$, while $y_i(\boldsymbol{x})$ is the outlier score of the faulty sample.

Since the number of faulty samples is limited, we augmented the separated data set with a parameter representing how many of the samples deviated from the normal situation. In the figure, we showed the results for three different choices of the

parameter (3, 6, and 12). In all cases, our approach is significantly better than the state-of-the-art. Note that the state-of-the-art method is based on expert knowledge, which means that our model is doing a better job with this metric than the best experts. This is a key advantage of the data-driven approach, which enables us to capture hidden patterns in the data.

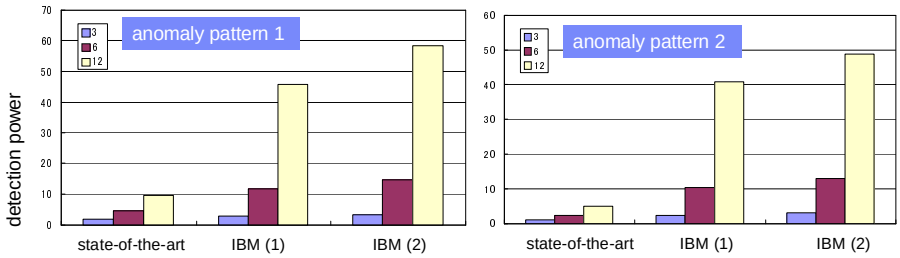

Fig. 3 Comparison of our approach with a state-of-the-art approach used by experts.

Finally, we show in Fig. 4 a screenshot of our fault detection method implemented on SPSS Modeler^{TM}. We see that a custom node is created in the Modeler window. By double-clicking the icon, we can edit parameters such as ρ in the model. The rich graphical user interface of the SPSS modeler provides users with very good usability.

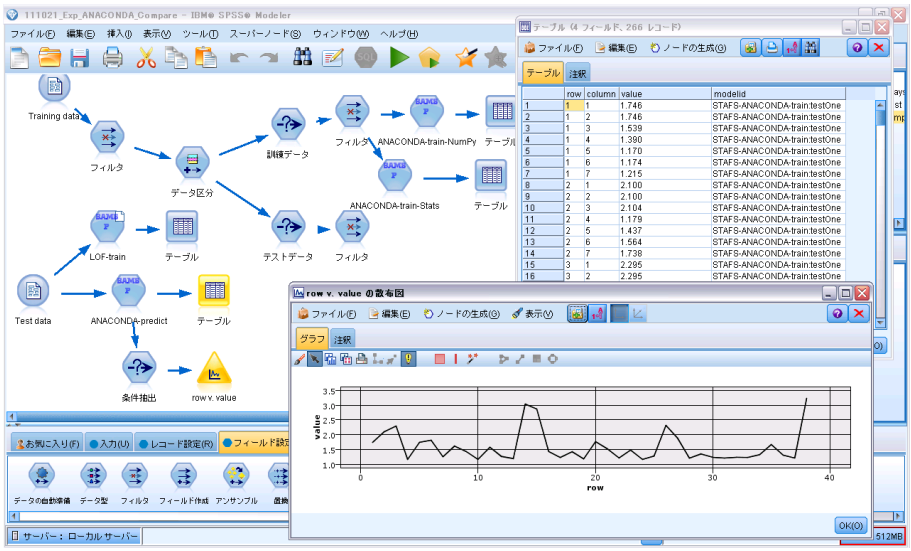

Fig. 4 Fault detection system implemented on SPSS Modeler^{TM}.

6 Summary

We have discussed how to formalize expert knowledge, which we believe is one of the key questions in service science. Based on the history of expert systems in AI, we suggested that three criteria, generalizability, learnability, and actionability, are critical for extracted rules to be useful. Then we pointed out that natural languages and their variants are not necessarily the only choice for knowledge representation, and the use of mathematical language provides better generalizability. We also pointed out that the data-driven approach of machine learning is useful in service science. Finally, we conducted a case study on condition-based maintenance in the Japanese railway industry, where our proposed statistical outlier detection method was demonstrated to be useful in early anomaly detection in wheel axles.

References

1. O. Banerjee, L. E. Ghaoui, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proc. Intl. Conf. Machine Learning*, pp. 89–96. Press, 2006.
2. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.
3. B. Buchanan and e. E. H. Shortliffe. *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1984.
4. A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.
5. S. Drake. *Discoveries and Opinions of Galileo*. Anchor, 1957, 2nd. edition, 1957.
6. J. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79, 2010.
7. J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.
8. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
9. T. Idé, A. C. Lozano, N. Abe, and Y. Liu. Proximity-based anomaly detection using sparse structure learning. In *Proc. of 2009 SIAM International Conference on Data Mining (SDM 09)*, pp. 97–108.
10. S. Nakazawa. ASP News, No.55, `http://www7b.biglobe.ne.jp/~asp/aspnews55.html`, (*in Japanese*), 1998.
11. P. S. Usama M. Fayyad, Gregory Piatetsky-Shapiro and e. Ramasamy Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.
12. S. L. Vargoa, P. P. Magliob, and M. A. Akakaa. On value and value co-creation: A service systems and service logic perspective. *European Management Journal*, 26(3):145–152, 2008.
13. D. A. Waterman. *A guide to expert systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1985.
14. Wikipedia. MYCIN, `http://en.wikipedia.org/wiki/Mycin`, 2012.
15. Wikipedia. Rail transport in Japan, `http://en.wikipedia.org/wiki/Rail_transport_in_Japan` , 2012.