Discovery of Damage Patterns in Fuel Cell and Earthquake Occurrence Patterns by Co-Occurring Cluster Mining

Ken-ichi Fukui[a], **Daiki Inaba**[a],[*], and **Masayuki Numao**[a]
[a]The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
[*]currently working at Ricoh Ltd.

Abstract

We have proposed a novel data mining method called co-occurring cluster mining (CCM) for mining patterns from a sequence of multidimensional event data. The CCM first generates cluster candidates and then test the candidates based on clustering in the data space as well as co-occurrence degree in the event sequence. In searching appropriate clusters associated with co-occurrence, the search space is reduced by obtaining a dendrogram from a hierarchical clustering as the clustering procedure. In this paper, we show the potential of CCM with following two applications: (1) damage patterns in fuel cell and (2) earthquake occurrence patterns. In the fuel cell application, given a sequence of acoustic emission events, which comprise of waveform signal data of damages to a fuel cell, the mechanical interactions between components of the fuel cell are inferred from the mined co-occurrence patterns. Similarly, in the application of earthquakes, the interactions between distant earthquakes are extracted as co-occurrence patterns from a hypocenter catalog.

Introduction

Data mining is essentially the inductive extraction of knowledge from observed data and is now widely applied in various fields because of its generality. In this study, we focus on a novel task that combines well-known clustering techniques and frequent pattern mining or association rule mining. Clustering(Everitt et al. 2011; Xu and Wunsch-II 2008) attempts to produce groups of similar objects within a so-called data space or feature space, which is typically represented by a multidimensional numerical vector. Frequent pattern or association rule mining(Agrawal and Srikant 1994; Han, Pei, and Yin 2000) attempts to extract and list frequently appearing item sets, wherein an item is typically identified via nominal variables.

In this study, given a sequence of events, where each event is represented by a multidimensional numerical vector (e.g., sequence of signal events, image events, and position events), the goal is *to find and list pairs of clusters that co-occur in a sequence*. This task may induce novel applications—e.g., identify weather change patterns from a sequence of satellite images, infer health change patterns from a sequence of medical inspection data, and infer mechanical interaction patterns between components from a sequence of sounds of damage.

A straightforward method to achieve the above task is to first generate clusters within the data space, and then extract frequent patterns or association rules from the sequence of clusters as items. For example, Honda *et al.* quantized by self-organizing map (SOM) images of cloud data obtained via satellite, and then applied association rule mining to extract climate change information(Honda et al. 2002). As another example, Yairi *et al.* extracted association rules regarding anomaly detection after clustering time-series data transmitted via satellite(Yairi et al. 2001).

In the two-step approach that these examples utilized, clusters may contain data points that do not contribute to a certain pattern in a sequence, or may not contain data points that contribute to a pattern at all. The contribution to a pattern can be justified by co-occurrence degree of the data points within the different clusters. The cluster ranges should be selected by co-occurrence degree to exclude non-contributing data points to a pattern and include contributing data points.

To solve the above problem, we have proposed a novel algorithm called *co-occurring cluster mining (CCM)*. The CCM generates cluster candidates and test the candidates based on clustering in the data space as well as co-occurrence degree in the sequence(Inaba et al. 2012). More specifically, CCM extracts and lists pairs of clusters that have high intra-cluster density in the data space and simultaneously high inter-cluster co-occurrence in the sequence of events.

Various works have been done in spatio-temporal data mining. Especially, space-time scan statistics(Kulldorff 2001) detects outbreak regions (clusters) in certain period based on statistical tests, for example detecting region(s) and period(s) of infection disease. However, space-time scan statistics does not extract co-occurrence of different regions, the purpose is to detect single region and its period.

In this paper, we added one more application of the CCM to show the generality of the methodology; (1) damage patterns in fuel cell, which is already written in Inaba *et al.*(Inaba et al. 2012) and (2) earthquake occurrence patterns, which is a new application. In the fuel cell application, from a sequence of acoustic emission events, which are comprised of waveform signal data derived from damages to a fuel cell, the mechanical interactions between components of the fuel cell can be inferred from the mined co-occurrence patterns, which were reasonable and confirmed by domain experts. Similarly, in the application of earthquake analysis, interactions between distant earthquakes are extracted as co-occurrence patterns from the 2011 Tohoku Earthquake in Japan.

Co-occurring Cluster Mining

Problem Definition

In this section, we define the characteristics of data that our current work focuses on, and then define requirements of the co-occurrence pattern.

Definition 1 (*event sequence*). Suppose data set $\mathcal{D}$ with N numerical event data points $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,v})$, $(i = 1, \cdots, N)$ in v-dimensional space are obtained in order $\boldsymbol{x}_1 \prec \cdots \prec \boldsymbol{x}_N$.

Definition 2 (*segment*). Suppose an event sequence is divided into segments. More specifically, let data set $\mathcal{D}$ be denoted by

$$\mathcal{D} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i][\boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_j] \cdots [\boldsymbol{x}_k, \cdots, \boldsymbol{x}_N],$$

where $i < j < k < N$ and "[·]" refers to a segment (similar to market basket analysis).

The segments above are measured in minutes, days, and so on; furthermore, the length of these segments need not be regular. Given the above, the extracted co-occurrence patterns must satisfy the following three requirements:

Requirement 1 (*co-occurrence*). For two sets composed of events $\mathbf{A}, \mathbf{B} \subset \mathcal{D}$ ($\mathbf{A} \cap \mathbf{B} = \emptyset$), the co-occurring ratio of $\mathbf{A}$ and $\mathbf{B}$ must be high. Co-occurrence can be evaluated by the Jaccard coefficient by counting the number of segments that contain $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{A}$ or $\mathbf{B}$.

Requirement 2 (*frequency*). The number of times in which $\mathbf{A}$ and $\mathbf{B}$ co-occur in an event sequence must be high. Such occurrence frequency can be evaluated, for example, by the support score by counting the number of segments that contain $\mathbf{A}$ and $\mathbf{B}$.

Requirement 3 (*similarity*). For two event sets $\mathbf{A}$ and $\mathbf{B}$, events $\boldsymbol{x} \in \mathbf{A}(\mathbf{B})$ must be similar. The within-cluster similarity can be evaluated by the sum of squares within clusters (SSW), the average distance among all data points in a cluster, etc.

Requirements 1 and 2 are derived from frequent pattern mining between event sets (clusters), whereas requirement 3 is derived from the clustering of events. Given the above, we define the co-occurring cluster and co-occurrence pattern as follows:

Definition 3 (*co-occurring cluster*). If two sets $\mathbf{A}, \mathbf{B} \subset \mathcal{D}$ satisfy the above three requirements, set $\mathbf{A}$ is a co-occurring cluster of $\mathbf{B}$ and vice versa.

Definition 4 (*co-occurrence pattern*). With co-occurring clusters $\mathbf{A}$ and $\mathbf{B}$, $\mathrm{P}(\mathbf{A}, \mathbf{B}) = \{\mathbf{A}, \mathbf{B} | \mathbf{A} \cap \mathbf{B} = \emptyset\}$ is called a co-occurrence pattern.

Evaluation Function

In this section, we define an evaluation function to search for co-occurrence patterns defined in the above section. We search pairs of clusters $\mathbf{A}, \mathbf{B} \subset \mathcal{D}$ that maximize the following evaluation function:

$$\mathcal{L}(\mathbf{A}, \mathbf{B}) = \mathcal{F}(\mathbf{A}, \mathbf{B})^{\alpha} \cdot \mathcal{G}(\mathbf{A}, \mathbf{B})^(1 - \alpha). \qquad (1)$$

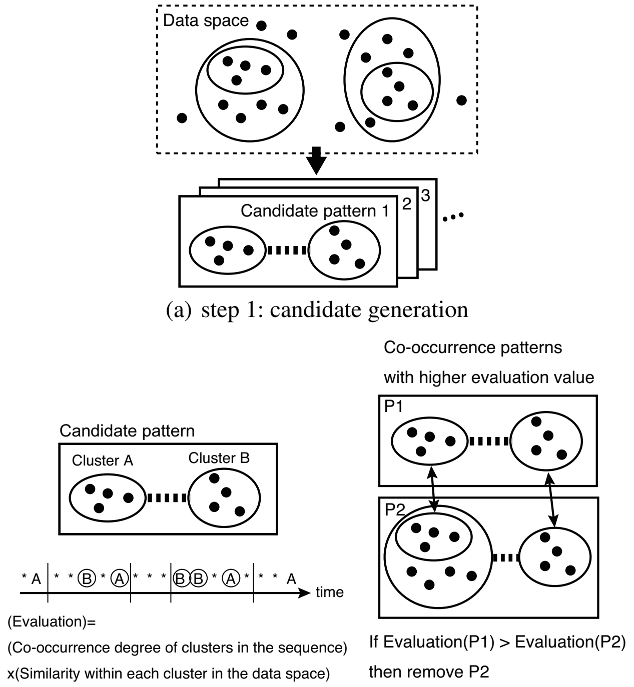
Function $\mathcal{F}(\mathbf{A}, \mathbf{B})$ evaluates the co-occurrence ratio for requirement 1. The higher the $\mathcal{F}(\mathbf{A}, \mathbf{B})$ value is, the higher the co-occurrence ratio. Note that because requirement 1 denotes the co-occurrence among many separated segments, co-occurrence in the short and sequential period must be excluded. Therefore, even if events from $\mathbf{A}$ and $\mathbf{B}$ co-occur several times in the same segment, this is considered only once. Function $\mathcal{G}(\mathbf{A}, \mathbf{B})$ denotes similarity within a cluster for requirement 3. The higher the $\mathcal{G}(\mathbf{A}, \mathbf{B})$ value is, the more dense the clusters are.

Evaluation function (1) is defined as the product of $\mathcal{F} \in [0, 1]$ and $\mathcal{G} \in [0, 1]$ to simultaneously satisfy the requirements of co-occurrence and similarity. α is a hyper-parameter to weight $\mathcal{F}$ or $\mathcal{G}$. In this work, $\alpha = 0.5$ is used for simplicity. In addition, requirement 2 for occurrence frequency can be satisfied using minimum support $Supp_{min}$ as a threshold.

CCM Algorithm

Our proposed method searches pairs of clusters that have high $\mathcal{L}(\mathbf{A}, \mathbf{B})$, utilizing aggregative hierarchical clustering (AHC) to reduce the search cost. To generate candidate clusters of $\mathbf{A}$ and $\mathbf{B}$, partition-based clustering techniques, such as k-means clustering, and probability-distribution-based clustering techniques, such as Gaussian mixture model clustering, need to be executed every time variables in the search are changed. Conversely, in AHC, once the merge process of clustering is obtained, co-occurrence patterns can be searched during the merge process. The other benefit of using AHC is to reduce the search space, although the degree of freedom for the cluster shape decreases. The number of sub-clusters from the dendrogram obtained by AHC—except for individual data points and whole data points as a cluster—is $N - 2$. Taking combinations of the sub-clusters as candidate patterns, the approximate computational complexity is $\mathrm{O}(N^2)$.

The pseudocode and conceptual diagram of the CCM algorithm are presented in Algorithm 1 and Fig. 1, respectively. The algorithm first generates possible sub-clusters from the dendrogram obtained by AHC in the data space. All combinations of sub-clusters can be candidate patterns (Step 1, Fig. 1(a)). Second, the algorithm evaluates each candidate pattern via function $\mathcal{L}$, which evaluates the co-occurrence degree of sub-clusters in the event sequence and similarity within each sub-cluster in the data space. If the evaluation value exceeds the minimum thresholds $\mathcal{L}_{min}$ and $Supp_{min}$, then these patterns are added to output pattern list $\mathcal{P}$ (Step 2, Fig. 1(b)). Here, the definition of support score (*l.* 6 of Algorithm 1) is $Supp(\mathbf{H}_i, \mathbf{H}_j) = count(\mathbf{H}_i \cap \mathbf{H}_j)/S$, where $count(\mathbf{H}_i)$ denotes the number of segments that contain event(s) with cluster label $\mathbf{H}_i$, and S is the number of


Data space
Candidate pattern 1
2
3
...
Co-occurrence patterns
with higher evaluation value
P1
P2
Candidate pattern
Cluster A
Cluster B
* A * * B * A * * * B B * A * * * A
time
(Evaluation)=
(Co-occurrence degree of clusters in the sequence)
x(Similarity within each cluster in the data space)
If Evaluation(P1) > Evaluation(P2)
then remove P2

(a) step 1: candidate generation

(b) step 2: evaluation of candidate patterns

(c) step 3: checking of inclusion between patterns

Figure 1: Conceptual diagram of co-occurring cluster mining algorithm

total segments. Third, the algorithm checks the inclusion between the patterns and removes patterns from $\mathcal{P}$ that have a lower evaluation score (Step 3, Fig. 1(c)). In the algorithm description, $\mathrm{P}_l \cap \mathrm{P}_m$ (*l.* 13 of Algorithm 1) means $\mathbf{A}_l \cap \mathbf{A}_m$ or $\mathbf{A}_l \cap \mathbf{B}_m$ or $\mathbf{B}_l \cap \mathbf{A}_m$ or $\mathbf{B}_l \cap \mathbf{B}_m$.

Application 1: Damage Patterns in Fuel Cell

Background

A fuel cell, especially a solid oxide fuel cell (SOFC), is a promising power generation device that produces electricity by direct chemical reaction; however, SOFC operates in harsh environments (i.e., high temperatures, oxidation, and reduction), and therefore, the reaction area is decreased by fracture damage, which reduces cell performance (Krishnamurthy and Sheldon 2004).

We have previously used a kernel self-organizing map (kernel SOM) approach to successfully model and visually understand the overview of the damage process from acoustic emission (AE) events(Fukui et al. 2011). Acoustic emission is an elastic wave (i.e., vibration, sound waves, including ultrasonic waves) produced by damage, such as cracks in the material, or by friction between materials. Depending on the “fracture mode” (i.e., opening or shear), the type of material, fracture energy, shear rate, and other factors, distinct AE waveforms are produced.

Little knowledge about the mechanical interactions of damages has been obtained until now. Hence, we applied

Algorithm 1 Co-occurring cluster mining algorithm

```
Input: event sequence with segments:
    $\mathcal{D} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i][\boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_j] \cdots [\boldsymbol{x}_k, \cdots, \boldsymbol{x}_N]$
    dendrogram by hierarchical clustering from $\mathcal{D}' = \{\boldsymbol{x}_k\}_{k=1}^N$: $\mathcal{HC}$
    minimum evaluation function value: $\mathcal{L}_{min}$
    minimum support value: $Supp_{min}$
Output: Co-occurrence patterns:
    $\mathcal{P} = \{\mathrm{P}_k(\mathbf{A}, \mathbf{B}) | \mathbf{A} \cap \mathbf{B} = \emptyset, \mathbf{A}, \mathbf{B} \subseteq \mathcal{D}\}$
 1: //Step 1: Generate sub-clusters
 2: Generate possible sub-clusters from $\mathcal{HC}$:
    $\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_{N-2} \subset \mathcal{HC}$;
 3: //Step 2: Evaluate candidate patterns
 4: Initialize $k \leftarrow 0$;
 5: for all combinations of H do
 6:     if $\mathbf{H}_i \cap \mathbf{H}_j = \emptyset$ and $\mathcal{L}(\mathbf{H}_i, \mathbf{H}_j) > \mathcal{L}_{min}$ and
        $Supp(\mathbf{H}_i, \mathbf{H}_j) > Supp_{min}$ then
 7:         $\mathrm{P}_k(\mathbf{A}, \mathbf{B}) \leftarrow \{\mathbf{H}_i, \mathbf{H}_j\}$;
 8:         $k \leftarrow k + 1$;
 9:     end if
10: end for
11: //Step 3: Eliminate co-occurrence patterns with inclusion relation
12: for all combinations of P do
13:     if $\mathrm{P}_l \cap \mathrm{P}_m \neq \emptyset$ then
14:         Remove $\mathrm{P}_i$ from $\mathcal{P}$ such that
            $i = \arg\min\{\mathcal{L}(\mathrm{P}_l), \mathcal{L}(\mathrm{P}_m)\}$;
15:     end if
16: end for
```

CCM to extract co-occurrence damage patterns that represent major mechanical interactions among components in SOFCs. Our experiments show that we can acquire novel knowledge—even for the SOFC experts—about damage mechanisms from co-occurrence damage patterns.

Experimental Conditions

A schematic of the apparatus used to perform SOFC damage testing is shown in Fig. 2. The test section was initially heated to 800°C to melt a soda glass ring, and was then gradually decreased to room temperature. Note that this damage evaluation test was designed to intentionally rupture the cells while lowering the temperature. Therefore, the knowledge obtained through this experiment is not directly available to actually run the SOFC; however, it is sufficient to demonstrate and confirm the reasonableness of our proposed method.

The AE measurement was performed using a wide-band piezoelectric transducer[1]. The AE transducer was attached to an outer $\mathrm{Al_2O_3}$ tube away from the heated section. The sampling rate was 1 MHz, and so the observable maximum frequency was 500 KHz.

Preprocessing

Running the SOFC for over 60 h, 1,429 AE events were extracted using the burst extraction method (Kleinberg 2002;

[1] PAC UT-1000: http://www.pacndt.com

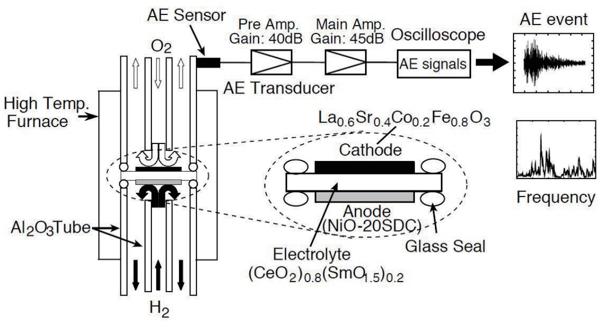


Figure 2: SOFC damage test apparatus

Fukui et al. 2011). In our research, the observed AE event sequence was divided into segments based on (Ohsawa 2002), assuming a sequence until a large-energy AE event occurs to be a chain of damage progression. Fig. 3 shows an example of the divided segments. These segments are used in CCM for calculating a co-occurrence. Note that because the damage process of SOFC is a complicated system, it is difficult to extract co-occurrence patterns considering the order of occurrence or the precise time intervals between the AE events. Therefore, we do not consider the order of occurrence of AE events in the same segment or the time intervals between the AE events.

More concretely, after the burst extraction method, an AE event sequence can be described by $\mathcal{D}_{ae} = \{\boldsymbol{e}_i | \boldsymbol{e}_1 \prec \cdots \prec \boldsymbol{e}_N\}_{i=1}^{N}$, $\boldsymbol{e}_i = (z_{i,1}, \cdots, z_{i,p})$, where $z_{i,j}$ is an observed value (mV) from an AE sensor. The approach is to first calculate the energy for all AE events, $E_1, E_2, \cdots, E_N$, where $E_i = \sum_j z_{i,j}^2$.

Next, divide the AE event sequence into segments $\boldsymbol{s} = [e_t, \cdots, e_{t+l}]$, each of which satisfies the following condition:

$$E_{t+i} \le E_\sigma \text{ and } E_{t+l} > E_\sigma \; (i = 0, \cdots, l-1), \tag{2}$$

where E_σ is an energy threshold; $E_\sigma = 1,500(\mathrm{mV}^2)$ is used in this paper. Just as in (Ohsawa 2002), there is no optimal threshold; however, the threshold was determined by finding a balance between the number of segments and the number of events contained within each segment. The AE event sequence was divided into 123 segments.

Also referring to our previous study (Fukui et al. 2011), each AE event was transformed into the frequency power spectrum $\boldsymbol{e^{sp}}_i = (sp_{i,1}, \cdots, sp_{i,v})$ by Fourier transform, where sp is a power of a certain frequency with v =4,000 discrete points in this study.

Design of the Evaluation Function

To extract a symmetric pattern, we used the Jaccard coefficient as $\mathcal{F}(\mathbf{A}, \mathbf{B})$ as follows:

$$\mathcal{F}(\mathbf{A}, \mathbf{B}) = \frac{count(\mathbf{A} \cap \mathbf{B})}{count(\mathbf{A} \cup \mathbf{B})}. \tag{3}$$

For $\mathcal{G}(\mathbf{A}, \mathbf{B})$, we cannot obtain the centroid of clusters, but can instead obtain the distance between the prototype vectors, with d_{ave} as the average distance among all pairs of prototype vectors in the cluster. Hence, $\mathcal{G}(\mathbf{A}, \mathbf{B})$ can be given by

$$\mathcal{G}(\mathbf{A}, \mathbf{B}) = 1 - \sqrt{\frac{d_{aveA} \cdot d_{aveB}}{d_{aveALL}^2}}, \tag{4}$$

$$d_{aveA} = \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbf{A}} \frac{d_{c(i),c(j)}}{N_A}, \tag{5}$$

$$d_{aveB} = \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbf{B}} \frac{d_{c(i),c(j)}}{N_B}, \tag{6}$$

$$d_{aveALL} = \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}} \frac{d_{c(i),c(j)}}{N_{\mathcal{D}}}, \tag{7}$$

where $i < j$, N_A is the number of combinations among the events in $\mathbf{A}$, and $d_{c(i),c(j)}$ denotes a distance between the best matching units for $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the topology space of the kernel SOM.

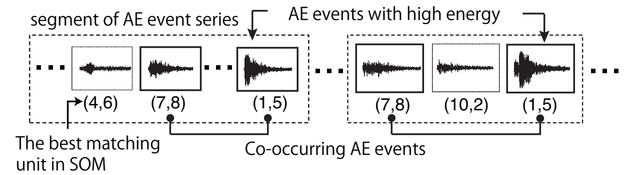


Figure 3: An example illustrating damage segments

Table 1: Average of evaluation scores of the extracted 100 patterns in different hierarchical clustering methods, including single linkage, complete linkage, group average, centroid, median, and Ward's method

	Single	Comp.	Group	Cent.	Med.	Ward's
Ave.	0.443	0.494	0.482	0.444	0.459	0.487

Results

The topology of the kernel SOM was set to a two-dimensional square grid, and the number of neurons is 15×15. The nodes in the SOM represent microclusters, and this number was sufficient for the interpretation, just as in (Fukui et al. 2011).

Table 1 shows the average values of the evaluation function using different hierarchical clustering methods. The values are averaged by the extracted 100 patterns when the minimum support $Supp_{min}$ is set to 0.04. The complete linkage method shows the best results. Therefore, the following results are obtained using the complete linkage method in the hierarchical clustering.

Tables 2 and 3 show the estimated interpretation of extracted damage patterns[2], which were provided by two of

[2] The interpretations were based on clustered AE events (waveforms and frequency spectrum), temperature at the time, and referring to the image of actual internal damages by an electron microscope after the operation.

Table 2: Major damage types corresponding to the results of the kernel SOM (Fukui et al. 2011)

region	damage type
(A)	squeaking of the members during heating
(B)	progression of the initial cracks
(C)	squeaking of the members followed by (B)
(D)	cracks in the electrolyte
(E)	cracks in the glass seal
(F)	cracks in and exfoliation of the electrode

Table 3: Number of extracted damage patterns in each damage type; the corresponding damage types are listed in Table 2, and the inter-region damage types are represented with "{,}"

pattern	number	pattern	number
(B)-(B)	2	(E)-(F)	5
(B)-(C)	3	(E)-{(A),(D)}	3
(B)-(D)	2	(E)-{(D),(E)}	1
(B)-(E)	2	(F)-{(A),(D)}	1
(C)-(C)	1	(F)-{(D),(E)}	1
(D)-(D)	1	{(A),(D)}-{(D),(E)}	1
(D)-(E)	1	{(D),(E)}-{(D),(E)}	1
(E)-(E)	4		

the SOFCs and fracture mechanics experts, and the number of extracted damage patterns by CCM, respectively. With parameters $\mathcal{L}_{min} = 0.47$ and $Supp_{min} = 0.04$[3], 29 patterns were extracted. The computational time when using the original 1,429 AE events was 888.7 s with an Intel Xeon CPU running at 2.66 GHz with 6 GB RAM. When using prototypes of the kernel SOM, the computational time in 225 (15×15) objects was reduced to 25.6 s.

Fig. 4 shows examples of the results of extracted damage patterns from the kernel SOM. The correspondence of the regions on the map to damage types is shown in Table 2. Each damage pattern is distinguished using different colors, and the typical waveforms and spectra of the damage types are shown.

Valid results based on the knowledge of SOFC experts: Damage pattern 1 in Fig. 4 is a co-occurring pattern of (B) "the progression of the initial cracks" and (D) "cracks in the electrolyte." Therefore, damage pattern 1 indicates that the progression of the initial cracks causes cracks in the electrolyte because of mechanical rationality. Table 3 indicates that the progression of the initial cracks co-occurs with various damages. We can interpret that the progression of the initial cracks largely affects various damage types.

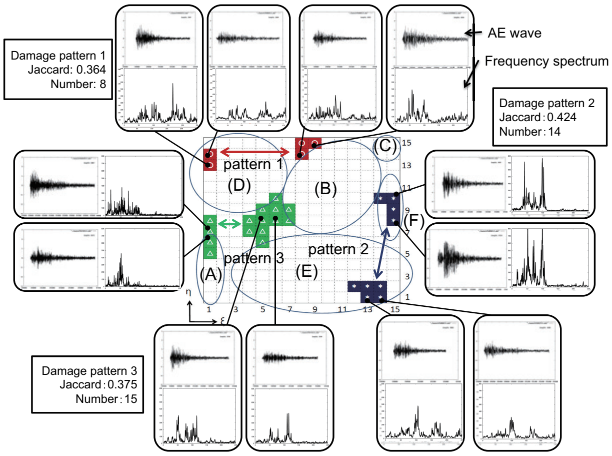


Figure 4: Examples of extracted damage patterns; the central map shows visualization results from the kernel SOM

Damage pattern 2 entails co-occurrence patterns (E) "cracks in the glass seal" and (F) "cracks in and exfoliation of the electrode." In particular, co-occurring cluster (E) is the latter period of cracks in the glass seal (from cluster change analysis in (Fukui et al. 2011)). The temperature is decreasing and the glass seal is congealed at the temperature of damage pattern 2. The glass seal and electrode are not directly connected, but it is supposed that the shrinking and transformation of the cell due to the coagulation of the glass seal produces the indirect mechanical effect.

Novel results even for the experts: According to Table 3, no damage patterns that include both regions (D) and (F) are extracted. Although the electrolyte and electrode are connected, damage patterns that include both were not extracted at all. This result was interesting to SOFC experts.

Furthermore, since damage pattern 3 exists in the inter-regions, it may contain novel damage types. Since these damages cause AE events containing high peaks in the low frequencies of the spectrum, the damages between regions (A) and (D) are estimated as "the exfoliation of the electrolyte," and those between regions (D) and (E) are estimated as "the exfoliation of the electrolyte or the glass seal." Damage pattern 3 has never been discovered from earlier research based only on the occurrence frequency of each AE event. Considering the co-occurrence relationship of AE events, damage pattern 3 is discovered and identified for the first time.

[3] These parameters were empirically determined after several trials with a comprehensive checking— i.e., scores of $\mathcal{F}$ and $\mathcal{G}$ should be well-balanced; the number of extracted patterns should be approximately less than 30 in order to manually analyze the patterns; when we set $\mathcal{L}_{min} < 0.47$, some co-occurring clusters are spread in very wide area which is intuitively a meaningless cluster; $Supp_{min} = 0.04$ means actually 5 times or more (segments) of co-occurrence, and we assumed patterns that have less than 5 times are low reliability.

Application 2: Earthquake Occurrence Patterns

Background

In this section, we describe our application of CCM to extract earthquake co-occurring patterns among areas afflicted by the 2011 Tohoku Earthquake in Japan, revealing affected areas and certain relationships between earthquakes. In seismology, it is said that earthquake activities are not predictable only from time, location, and magnitude, because they are highly sensitive and nonlinearly dependent (Geller et al. 1997); however, the seismic activities represent, in part, the internal state of the Earth's crust at the time. Thus, the analysis of such activities advances the understanding of the seismic occurrence mechanism.

Numerous studies utilize data mining for the analysis of seismic activities, including, for example, density-based clustering (Lei 2010) and fuzzy clustering (Ansari, Noorzad, and Zafarani 2009), which are used to find earthquake hot spots. Lee *et al.* (Lee, Han, and Chi 2009) used quantitative association rule mining and found relationships between features such as depth and magnitude, and location and frequency. Martínez-Álvarez *et al.* (Martínez-Álvarez et al. 2011) utilized a quantitative association rule and regression to investigate earthquake prediction in a specific area.

From the perspective of co-occurrence of earthquakes, to the best of our knowledge, Ohsawa's study (Ohsawa 2002) is the only one that extracted frequent co-occurrence patterns among active faults by KeyGraph, which was originally designed as a keyword-extraction method. His research considered seismic events that occurred in all Japanese islands from 1985 to 1992; earthquakes that occurred off the coast were excluded. This study also utilized a two-step approach, wherein hypocenters were categorized using the nearest active fault names rather than via clustering.

Data Preprocessing

We applied CCM to the hypocenter catalog data recorded for the calendar year 2011, as released by the Japan Meteorological Agency (JMA) through the Japan Meteorological Business Support Center[4]. Each event has an origin time (JST), a hypocenter (latitude, longitude, and depth), a magnitude, and a hypocenter area name. Events with the maximum seismic intensity greater than three were recorded in the catalog; 738 seismic events were recorded in that period. We used only latitude and longitude as the attributes for merging clusters because there are only a few differences among depths in the same areas.

Regarding segmentation of the seismic event sequence, much the same as in the fuel cell application, the key idea is based on Ohsawa's study (Ohsawa 2002)—i.e., the segment division was performed utilizing the magnitude of the seismic events. When a large energetic event occurs, the structure of the Earth's inner crust changes, and the seismic process will transit to another condition; however, in the Tohoku Earthquake, quakes greater than M6.0 occurred near the mainshock[5] at short time intervals. Therefore, the length of the segments decreased immediately after the mainshock. We then introduced time constraints for segment length to eliminate very short segments.

On the basis of the above ideas, a seismic event sequence can be divided into segment $\boldsymbol{s} = [\boldsymbol{x}_s, \cdots, \boldsymbol{x}_{s+l}]$ by the following two conditions:

$$E_{s+i} \leq E_\sigma \text{ and } E_{s+l} > E_\sigma \ (i = 0, \cdots, l-1), \tag{8}$$

$$t_{s+l} - t_s > t_\sigma, \tag{9}$$

where E_i and t_i denote corresponding magnitude and origin time of earthquake event $\boldsymbol{x}_i$, respectively; E_σ is set to M6.0 and t_σ is one hour. With these conditions, we obtained 59 segments with an average segment length of 12.5 events, which is approximately the same as that seen in Ohsawa's study (Ohsawa 2002).

Design of the Evaluation Function

For $\mathcal{F}(\mathbf{A}, \mathbf{B})$, the Jaccard coefficient (eq. (3)), same as in the fuel cell application, was used. Unlike the fuel cell application, $\mathcal{G}(\mathbf{A}, \mathbf{B})$ is defined by the following function:

$$\mathcal{G}(\mathbf{A}, \mathbf{B}) = \exp\left(-\frac{SSW(\mathbf{A})^2 + SSW(\mathbf{B})^2}{2\sigma^2}\right). \tag{10}$$

Here, we apply an exponent because of a bias of SSW in this dataset, where σ is a parameter to control correction of the bias. We used $\sigma = 0.05$ in this dataset, which maximizes the average evaluation function values.

Results

Validation of the Extracted Patterns The functional parameters were set to $\mathcal{L}_{min} = 0.60$ and $Supp_{min} = 0.08$[6], and Ward's method was used for hierarchical clustering. With these parameters, 15 earthquake co-occurrence patterns were obtained via CCM. All extracted patterns with evaluation scores are listed in Table 4. The symbols within parentheses indicate certain geometric areas corresponding to co-occurring clusters, and the inclusion of a prime indicates a hierarchical relation of clusters, for example, A$'$ $\subset$ A.

Here, the suspected patterns may also be extracted. If both events $\mathbf{A}$ and $\mathbf{B}$ occur in most segments, the Jaccard coefficient appears to be high since it does not consider the case in which neither event $\mathbf{A}$ nor $\mathbf{B}$ occurs. We checked these suspected patterns by using Fisher's exact probability test. For example, the component of $[\mathrm{A}_+, \mathrm{B}_+]$ (both A and B are positive) in the contingency table was determined by calculating the number of segments in which both events $\mathbf{A}$ and $\mathbf{B}$ occur. If the p-value was greater than 0.05, we regarded the pattern as a suspected pattern. As a result, all extracted patterns satisfied p-value < 0.05, as shown in Table 4, which signifies that no patterns were obtained simply by chance.

Fig. 5(a) shows representative seismic patterns plotted with a geographic information system. We obtained distant

[4] http://www.jmbsc.co.jp

[5] A mainshock is the largest earthquake in a series of related earthquakes.

[6] These thresholds were also determined by the same way of comprehensive checking in the fuel cell application.

Table 4: Scores of the extracted seismic patterns; pattern IDs correspond to Fig. 7

Pattern	$\mathcal{L}$	$\mathcal{F}$	$\mathcal{G}$	p-value	number
$\mathrm{P}_1(\mathrm{B}'', \mathrm{D})$	0.91	0.83	0.99	1.19e-06	5
$\mathrm{P}_2(\mathrm{C}, \mathrm{J})$	0.61	0.38	0.96	1.59e-03	5
$\mathrm{P}_3(\mathrm{M}', \mathrm{F})$	0.63	0.42	0.97	8.99e-04	5
$\mathrm{P}_4(\mathrm{N}, \mathrm{A}')$	0.77	0.63	0.94	2.44e-05	5
$\mathrm{P}_5(\mathrm{A}, \mathrm{D})$	0.74	0.56	0.98	2.52e-05	5
$\mathrm{P}_6(\mathrm{A}, \mathrm{B}')$	0.70	0.50	0.98	1.42e-04	5
$\mathrm{P}_7(\mathrm{E}', \mathrm{D})$	0.76	0.63	0.91	1.12e-05	5
$\mathrm{P}_8(\mathrm{E}', \mathrm{B}'')$	0.71	0.56	0.91	6.40e-05	5
$\mathrm{P}_9(\mathrm{O}, \mathrm{L}')$	0.60	0.42	0.87	1.16e-03	5
$\mathrm{P}_{10}(\mathrm{G}, \mathrm{B}')$	0.63	0.42	0.94	8.99e-04	5
$\mathrm{P}_{11}(\mathrm{K}, \mathrm{I})$	0.63	0.45	0.87	4.65e-04	5
$\mathrm{P}_{12}(\mathrm{E}, \mathrm{A})$	0.65	0.50	0.86	7.00e-05	7
$\mathrm{P}_{13}(\mathrm{E}, \mathrm{G})$	0.62	0.46	0.82	1.30e-04	6
$\mathrm{P}_{14}(\mathrm{H}, \mathrm{M})$	0.62	0.50	0.78	1.42e-04	5
$\mathrm{P}_{15}(\mathrm{L}', \mathrm{B})$	0.66	0.52	0.82	1.50e-05	11

seismic patterns, such as P_2, and patterns between inland and shore events, such as P_9 and P_{11}. Such patterns are difficult to extract from only the distribution of hypocenters (Fig. 5(b)).

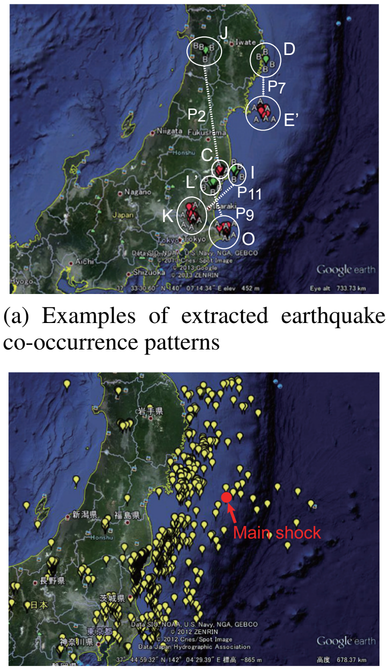
Comparison to the Two-Step Method

Fig. 6 shows a box plot comparing $\mathcal{F}(\mathbf{A}, \mathbf{B})$ and $\mathcal{G}(\mathbf{A}, \mathbf{B})$ for the 15 extracted patterns by CCM and the two-step method. The two-step method used hypocenter area names as clusters and extracted frequent item sets based on the Jaccard coefficient. CCM clearly provided a higher co-occurrence ratio by $\mathcal{F}$ and cluster compactness by $\mathcal{G}$ than those of the two-step method, especially in cluster compactness. Therefore, we conclude that CCM can determine cluster ranges that are related to a co-occurrence better than the two-step method.

Seismic Pattern Network The extracted seismic patterns can be connected by utilizing the hierarchical relation of clusters, as shown in Fig. 7. There are some regions in which co-occurrence patterns exist between more than three areas. For example, the southeastern area of Fukushima Prefecture (A and B$''$), off the coast of Miyagi Prefecture (E and E$'$), and off the coast of Iwate Prefecture (D) form a complete graph. These areas can be highly seismically related. Off the coast of Iwate Prefecture (D) is discriminative; even though only one area was extracted, this area is a co-occurrence cluster of three patterns P_1, P_5, and P_7, indicating that (D) is a highly influential area. We can also interpret from the network that the northern area of the Ibaraki Prefecture (M$'$, M, L$'$, and L) is a highly influential area that has relations in the four patterns P_3, P_9, P_{14}, and P_{15}.

Conclusion

We described CCM as a novel data mining approach for extracting pairs of clusters corresponding to co-occurrences in a sequence of events. The CCM algorithm searches clusters



(a) Examples of extracted earthquake co-occurrence patterns

(b) All earthquake events in 2011 with maximum seismic intensity greater than three

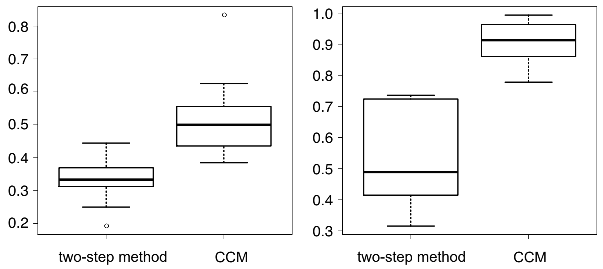
Figure 5: Distribution of hypocenters

that are dense in the data space and simultaneously co-occur in the sequence of events. The co-occurrence patterns are searched within the dendrogram obtained by a hierarchical clustering, which reduces the search space, and are extracted by maximizing the evaluation function of both similarity within clusters and co-occurrence of clusters.

In the application of a fuel cell, from a sequence of acoustic emission events of damage to the cell, we demonstrated that CCM can reveal mechanical interactions among components of the fuel cell. Next, in the application of earthquake analysis, from a sequence of seismic events of hypocenters, interactions among seismic activities can be obtained via CCM. Some seismic patterns were geographically distant or between island and shore; also, highly influential areas were identified; however, verification of the extracted patterns on seismological adequateness is difficult, but important for our future work. These applications show the generality of CCM, and CCM has a potential to open new analytics for multidimensional event sequences to reveal interactions among such events.

Acknowledgment

This study was supported by JSPS KAKENHI Grant Number 24650068. We also thank Prof. Kazuhisa Sato and Prof. Junichiro Mizusaki in Tohoku University for providing the experimental data of the fuel cell and discussion on the interpretation of our results.



(a) $\mathcal{F}(\mathbf{A}, \mathbf{B})$ (b) $\mathcal{G}(\mathbf{A}, \mathbf{B})$

Figure 6: Box plot of evaluation values for the extracted seismic patterns

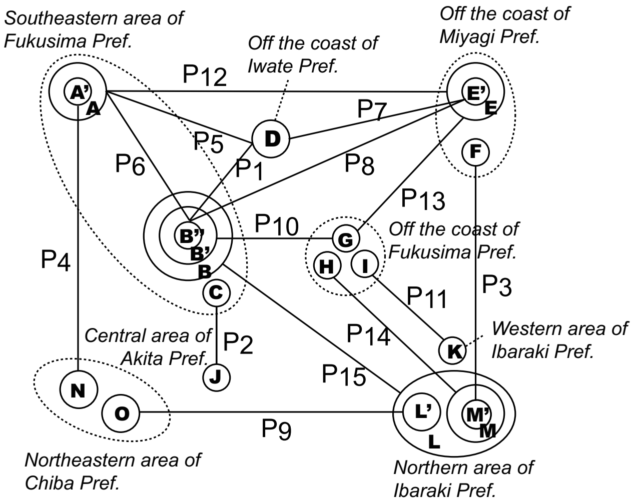


Figure 7: Network of co-occurrence relationships among all extracted patterns

References

Agrawal, R., and Srikant, R. 1994. Fast algorithms for mining association rules. In *Proc. of 20th International Conference on Very Large Databases (ICVLD)*, 487–499.

Ansari, A.; Noorzad, A.; and Zafarani, H. 2009. Clustering analysis of the seismic catalog of Iran. *Computers & Geosciences* 35:475–486.

Everitt, B. S.; Landau, S.; Leese, M.; and Stahl, D. 2011. *Cluster Analysis, 5th Edition*. Wiley.

Fukui, K.; Akasaki, S.; Sato, K.; Mizusaki, J.; Moriyama, K.; Kurihara, S.; and Numao, M. 2011. Visualization of damage progress in solid oxide fuel cells. *Journal of Environment and Engineering* 6(3):499–511.

Geller, R. J.; Jackson, D. D.; Kagan, Y. Y.; and Mulargia, F. 1997. Earthquakes cannot be predicted. *Science* 275(5306):1616.

Han, J.; Pei, J.; and Yin, Y. 2000. Mining frequent patterns without candidate generation. In *Proc. of the ACM SIGMOD Conf. on Management of Data*, 1–12.

Honda, R.; Wang, S.; Kikuchi, T.; and Konishi, O. 2002. Mining of objects from time-series images and its application to satellite weather imagery. *Journal of Intelligent Information Science* 19(1):79–93.

Inaba, D.; Fukui, K.; Sato, K.; Mizusaki, J.; and Numao, M. 2012. Co-occurring cluster mining for damage patterns analysis of a fuel cell. In *Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-12)*, volume LNAI 7301, 49–60.

Kleinberg, J. 2002. Bursty and hierarchical structure in streams. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, 91–101.

Krishnamurthy, R., and Sheldon, B. W. 2004. Stresses due to oxygen potential gradients in non-stoichiometric oxides. *Journal of Acta Materialia* 52:1807–1822.

Kulldorff, M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A* 164:61–72.

Lee, J. A.; Han, J. G.; and Chi, K. H. 2009. Mining quantitative association rule of earthquake data. In *International Conference on Convergence and Hybrid Information Technology*, 349–352.

Lei, L. 2010. Identify earthquake hot spots with 3-dimensional density-based clustering analysis. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010)*, 530–533.

Martínez-Álvarez, F.; Troncoso, A.; Morales-Esteban, A.; and Riquelme, J. C. 2011. Computational intelligence techniques for predicting earthquakes. In *Hybrid Artificial Intelligence Systems*, 287–294.

Ohsawa, Y. 2002. Keygraph as risk explorer in earthquake-sequence. *Journal of Contingencies and Crisis Management* 10(3):119–128.

Xu, R., and Wunsch-II, D. C. 2008. *CLUSTERING*. IEEE Press Series on Computational Intelligence.

Yairi, T.; Ishihama, N.; Kato, Y.; Hori, K.; and Nakasuka, S. 2001. Anomaly detection method for spacecrafts based on association rule mining. *Journal of Space Technology and Science* 17(1):1–10.