A methodology based on symbolic data analysis for structural damage assessment

Alexandre Cury, Christian Crémona
*University Paris-East, Laboratoire Central des Ponts et Chaussées, Paris, France*

Edwin Diday
*Paris IX Dauphine University, Paris, France*

ABSTRACT: This paper proposes the use of Symbolic Data Analysis (SDA) methods applied to damage classification. The methods are based on the knowledge of the modal parameters as well as to the raw knowledge of the structure’s acceleration and/or displacement. Some basic concepts of SDA and its application to a real case are introduced: this includes hierarchy-divisive methods, dynamic clustering and pyramid schemes for data classification based on raw measurements or structural modal parameters. Results are presented in order to show the efficiency of the described methodology.

1 INTRODUCTION

Studies related to early damage detection is of special concern for civil engineering structures. It is common knowledge that if a damage process is not identified in time, structural systems may have serious safety and economic consequences. Traditional methods of damage detection and health monitoring are often based on the variation of structural vibration characteristics, i.e. natural frequencies, damping ratios and mode shapes. These modal parameters are directly affected by changes in the physical properties of the structure including its mass and stiffness. Nevertheless, modal parameters identification is a sort of filtering process, leading to a loss of information compared to the raw data. This compression process can erase any small changes due to a structural modification. In turn, using raw dynamic measurements (especially if high sampling frequencies are used) leads to the storage of large set of data. Dynamic measurements can easily contain over thousands of values making an analysis process extensive and prohibitive. Nevertheless, several damage detection methods exist in the literature based on signature principles, but they usually fail when making them practical. In this sense, despite the current processing power of computers, the necessary computational effort to manipulate large data sets remains a problem. Furthermore, and this is certainly the major drawback when using modal parameters, is that modal components are essentially describing an equivalent linear behaviour, a feature which may be not exact for the analysis of specific degraded systems.

*Data mining* is the process of extracting hidden patterns from data. As more data is gathered in monitoring, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size. However, while it can be used to uncover hidden patterns in data that has been collected, obviously it can neither uncover patterns which are not already present in the data, nor can it uncover patterns in data that has not been collected. In order to deal with this issue, it is important to recall which types of data can be employed and manipulated in data mining (Billard et al, 2006) such as:

- A single quantitative value, e.g. `height(w) = 3.5;` where `w` is an individual;
- A single categorical value. For example: `town(w) = London;`
• Interval-valued data. For instance: weight \( w = [20, 180] \) which means that the weight of \( w \) varies in the interval \([20, 180]\);
• Multi-valued categorical data. For example: price\( (w) = \{\text{high, average, low}\} \) meaning that the general price of a product \( w \) may be high, average or low;
• Modal multi-valued (a histogram function). For instance: height \( (w) = \{[0, 1.20] (0.225); [1.20, 1.50] (0.321); [1.50, 1.80] (0.335); [1.80, 2.10] (0.119)\} \) meaning that 22.5% of the population \( w \) has a height varying in the interval \([0, 1.20]\).

This richer type of data is called symbolic data and it allows representing the variability and uncertainty present in each variable. The development of new methods of data analysis suitable for treating this type of data is the aim of Symbolic Data Analysis (SDA). Most of the currently developed techniques in symbolic data analysis are extensions of statistical methods. (Diday et al, 2008) certify the growth of data of symbolic nature and alert to the necessity of developing new statistical methodologies for the treatment of such information. In general, SDA provides suitable tools for managing complex, aggregated, relational, and higher-level data. The methodological issues under development generalize the classical data analysis techniques, like visualization, factorial techniques, decision tree, discrimination, regression as well as classification and clustering methods. Previous works in the literature introduced the use of SDA applied to data representing structural temperature variation in time. (Diday et al, 2001) showed promising results and important remarks for the perspectives of the present study.

The purpose of this paper is to present some principles of symbolic data analysis and, more specifically, the use of clustering methods applied to structural damage assessment. The idea consists in using different classification procedures in order to have insights about structural health conditions. In other words, the SDA is applied to either the vibration data (signals) obtained through dynamic tests or the modal parameters in order to infer if a damage process is in progress or if it has already occurred in the structure. In order to show the potentialities of the proposed methodology, several results regarding experimental tests performed on a real case, a rail bridge located in France on the high speed line between Paris and Lyon. Vibration measurements were obtained under three different conditions: before, during and after a structural modification process.

2 SYMBOLIC DATA OVERVIEW

In this section, a few principles of SDA as well as some specific terms often used in the analyses are presented. Moreover, the differences between classical and symbolic data as well as the methodology applied in order to transform the first one into the latter one is explained.

2.1 Classical data versus symbolic data

It is possible to argue that while the classical analysis is focused on studying individuals, symbolic analysis will deal with concepts that represent a richer and less specific type of data. Table 1 presents a few examples.

<table>
<thead>
<tr>
<th>Table 1 – Comparison between the units of study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classical INDIVIDUALS</strong></td>
</tr>
<tr>
<td>Player</td>
</tr>
<tr>
<td>Roses, orchids</td>
</tr>
<tr>
<td>Pigeons, ducks</td>
</tr>
<tr>
<td>Measured values</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 - Horse breeds</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breed</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
The first step in symbolic data analysis is to gather and describe these concepts. This can be achieved by using variables that allow not only representing but also explaining those concepts. Tables 2 and 3 show a comparison between two data sets containing different examples of application data.

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensor 1 (m.s(^2))</th>
<th>Sensor 9 (m.s(^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[0.03, 0.05]</td>
<td>[0.41, 0.62]</td>
</tr>
<tr>
<td>2</td>
<td>[0.07, 0.08]</td>
<td>[0.55, 0.77]</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>20</td>
<td>[0.05, 0.05]</td>
<td>[0.38, 0.49]</td>
</tr>
</tbody>
</table>

In Table 2 the concept “Horse breed” is studied and classified according to two variables: “Minimum Weight” and “Color”. It gathers a total of 5000 individuals (horses) into 10 concepts (breeds). The variable “Minimum Weight” is defined as an interval-valued data. For Breed 1, for example, the minimum weight varies between 186 kg and 209 kg. The variable “Color” is defined as a multi-valued qualitative data. For Breed 1, for instance, the colour can be either white or grey. In Table 3 the concept “Tests” is studied and classified according to two variables: “Sensor 1” and “Sensor 9”. It gathers a total of 80000 individuals (measured values) into 20 concepts (tests). Both variables “Sensor 1” and “Sensor 9” are defined as an interval-valued data. For Test 1, for example, the values measured by “Sensor 1” vary between 0.03 m.s\(^2\) and 0.05 m.s\(^2\). Table 3 actually represents the studied case in this paper and more details will be presented further.

2.2 Transforming classical data into symbolic data

Fig. 1 illustrates two hypothetical dynamic tests. Both tests include the measurements at two sensors for a total of 5000 acceleration values each.

![Figure 1 – Example of two experimental tests](image)

At this point in the paper, SDA will be considered by taking into account measured values for individuals, sensors for variables and tests for concepts. There are several ways to transform classical data into symbolic data. Let us consider, for instance, a signal \( X \) containing 5000 acceleration values measured by one single sensor. This signal can be represented by:

- A \( n \)-class histogram: \( X = \{(1(0.0025), 2(0.0721), 3(0.8546), 4(0.0626),..., n(0.0082))\};
- A min/max interval: \( X = [-0.025; 0.025]; \)
- An interquartile interval: \( X = [-0.012; 0.015]. \)

In this paper, two types of symbolic data are used in the SDA process: 20-class histogram-valued and interquartile intervals. They were chosen as they are properly describing the original classical data. Nevertheless, to use such descriptions is a filtering process since they reduce the amount of data to a smaller set. In turn, we can expect that non-linear features, if existing, are
not eliminating from this process. In order to obtain the first type of data (histogram-valued), it is necessary to perform a projection of each measured value to the Y-axis of coordinates. Fig. 2 shows how to construct a histogram for each sensor from each dynamic test. The divisions of a histogram are denominated classes.

![Figure 2 – Transforming classical data into symbolic data](image)

3 PRINCIPLES OF SYMBOLIC CLUSTERING METHODS

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics (Bock et al, 2001). A clustering procedure can be defined as a way of classifying a number of objects into different groups. More precisely, it can be described as the partitioning of a data set into subsets (clusters), so that the data in each subset share some common properties. In other words, for an appropriate clustering, it is necessary to minimize the within-cluster variation in order to obtain the most homogeneous clusters as possible, and to maximize the between-cluster variation in order to obtain the most dissimilar clusters among each other. In order to define these clusters and determine the proximity (or similarity) among the concepts, it is necessary to define suitable dissimilarity measures, which will be described in the next section.

3.1 Dissimilarity Measures and Within Variations

The formation of clusters \((C^1, \ldots, C^r)\) is governed by dissimilarity measures. These measures can supply numerical values in order to show the distance between two objects. In a common sense, the lower these values are, the more similar the objects are and thus, they are gathered in the same cluster. Conversely, the objects allocated into different clusters are the ones which have greater distances between them. (Cole et al, 1998) asserts that a small distance between two objects indicates high similarity. Thus, a distance measure can be used to quantify similarities as well as dissimilarities. Dissimilarity measures can take a variety of forms and some applications might require specific ones (Malerba et al, 2001 and Malerba et al, 2002). An important step for any clustering method is to select a convenient dissimilarity measure. These measures will influence the shape of the clusters. For example, a given object \(A\) may be close to a given object \(B\) according to one distance but distant according to another distance. In this work, for the interval-valued data, the Hausdorff distance is used. Let us consider two tests \(T_i\) and \(T_j\) which have their interquartile values represented by \((T_{i,\text{inf}}, T_{i,\text{sup}})\) and \((T_{j,\text{inf}}, T_{j,\text{sup}})\), respectively. The distance measure \(\phi(T_i, T_j)\) can be calculated as shown in Eq. (1):

\[
\phi(T_i, T_j) = \left( \frac{1}{n_{\text{sensors}}} \sum_{p=1}^{n_{\text{sensors}}} \left( \max\{ |T_{i,\text{inf}} - T_{j,\text{inf}}|, |T_{i,\text{sup}} - T_{j,\text{sup}}| \} \right)^2 \right)^{1/2}
\]

where \(n_{\text{sensors}}\) represents the number of sensors considered.

As previously mentioned, in order to have the most homogeneous clusters as possible, it is
necessary to minimize their within-variations $I$. This can be achieved by minimizing the total within-cluster variation $W$. The within variation of a cluster $C^k$ containing tests denoted by $T_i$ can be calculated as:

$$ I(C^k) = \frac{1}{m_k} \sum_{i=1}^{n_k} \sum_{j=i+1}^{n_k} \phi^2(T_i, T_j) $$

where $n$ is the total number of tests, $n_k$ is the number of tests of cluster $C^k$ and $\phi(T_i, T_j)$ are the elements of the dissimilarity matrix and can be calculated according to Eq. (1), for example.

The total within-cluster variation $W$ can be evaluated as the sum of all clusters within-variations:

$$ W = \sum_{k=1}^{r} I(C^k), $$

where $r$ is the number of clusters.

3.2 Clustering methods

The clustering simulations were carried out using the software SODAS (Symbolic Official Data Analysis System), developed under the project ASSO (Analysis System of Symbolic Official Data). There are several clustering methods present in the literature. In this paper, three methods will be analyzed:

- **Hierarchy-Divisive** clustering which is based on successive top-down divisions;
- **Dynamic Clouds** which consists in gathering the nearest tests according to a specific iterative algorithm;
- **Hierarchy-Agglomerative**, based on a down-top agglomeration process.

3.2.1 Hierarchy-Divisive

Divisive clustering is a top-down clustering process that starts with the entire data set as one cluster $C$ and then proceeds downward through as many levels as necessary to produce the hierarchy $Y = (C^1, C^2, \ldots, C^r)$. Theoretically, it is possible to have a number of clusters $r$ consisting of a single test. In practice, however, the clustering process stops at an earlier stage. Typically, a maximum value for $r$ is specified. Let us assume a single cluster $C$ containing all tests denoted by $(T_1, T_2, \ldots, T_r)$ and proceed by dividing this cluster into two clusters ($C^1$ and $C^2$). This procedure is carried out by determining which tests satisfy or not some logical criterion $q(.)$ (True or False). For instance, $q_1$: Is Sensor $4 \leq 0.05$ m.s$^{-2}$? If “True” (tests which satisfy the criterion), then they fall into the cluster $C^1$, if “False”, then those tests fall into $C^2$. In this example, “Sensor 4” and the value “0.05 m.s$^{-2}$” represent the variable and the cut-value $c$ which minimize the total-within variation $W$ of clusters $C^1$ and $C^2$. In other words, “Sensor 4” represents the most discriminant variable.

Fig. 3 shows a simplified scheme for the first partitioning procedure. Briefly, the clustering process consists of two iterative procedures: the first one corresponds to an iteration considering the number of sensors used ($n_{sensor}$) and the second one, an iteration considering the number of cut-values ($n_{cut}$). For further details, the reference (Billard et al, 2006) is advised. For each set “(sensor, cut-value)”, both within- and between-cluster variations are evaluated. This procedure continues until the absolute minimal value for the sum $W$ of within variations is obtained. Finally, the partitioning process will go on by taking into account the clusters corresponding to the minimal $W$ value. In that sense, the clustering process will continue by partitioning the clusters $C^1$ and $C^2$.

3.2.2 Dynamic Clouds

This clustering method is based on a generalization of the classical dynamical clusters method ($K$-means algorithm) (Lechevallier, 1974). This method consists in minimizing a general optimized criterion that measures the adequacy between the partition and the representation of the clusters, denoted prototype (Diday, 1971 and Celeux et al, 1989). A prototype is a symbolic de-
scription model to represent a cluster or, in other words, the “average” concept of a cluster. It is used as a reference to calculate the distances among the concept and to define each cluster. The algorithm starts with a set of $k$ random prototypes and iteratively applies an allocation step to place each concept in the cluster where the proximity between concept and prototype is minimal. In the sequence, a representation step is performed where the prototypes are updated according to the allocation step results. This is accomplished by computing and storing the total sum of the distances between concepts and the prototype in one cluster. The new cluster prototype will be the one which minimizes this sum. These two steps are repeated until the adequacy criterion reaches a stationary value (or at least when reaching a maximum number of iterates). In general, the dynamic cluster algorithm converges in a few iterations. In order to improve the quality of the clustering procedure, the algorithm is executed a different number of times with different initial partitions, so the best configuration is chosen as result. Fig. 4 shows a simplified scheme of this clustering method.

```
input : Cluster $C(T_1, T_2, \ldots, T_n)$
output: Clusters $C^1$ and $C^2$ with minimal $W$
1 for $i \leftarrow 1$ to $n_{sensors}$ do
2 for $j \leftarrow 1$ to $n_{prot}$ do
3 if Sensor $i \leq c(j)$ then
4 $C^1_{ij} \leftarrow \text{tests “True”;}
5 else
6 $C^2_{ij} \leftarrow \text{tests “False”;}
7 $W_{ij} \leftarrow I(C^1_{ij}) + I(C^2_{ij});$
8 $(i_{opt}, j_{opt}) \leftarrow \text{Argmin}_{ij} W_{ij};$
9 end
10 end
11 Repeat procedure with $C^1_{(i_{opt}, j_{opt})}$ and $C^2_{(i_{opt}, j_{opt})}$.
```

Figure 3 – Hierarchy-divisive algorithm.

Figure 4 – Scheme representing the dynamic clouds algorithm.
3.2.3 Hierarchy-Aggglomerative

Hierarchy-Aggglomerative is a bottom-up clustering process, so the principles of cluster construction outlined for the Hierarchy-Divisive clustering apply in reverse (Brito, 1998). Instead of starting with the complete set of tests and partitioning the cluster C into two subclusters at each stage (C^i and C^j), the process starts with r clusters containing one single test, and proceeds by merging two subclusters (C^i and C^j, say) into one new cluster C. The subclusters are merged according to similar criteria of minimizing the within-cluster variation and maximizing the between-cluster variation previously explained. In this sense, the clustering process starts by gathering the closest tests according to “generality” measures. More details can be found in (Diday et al, 2008). This clustering method can also reflect a measure of proximity between clusters by means of the “difference in height” between them. Fig. 5 illustrates a hierarchy-agglomerative clustering procedure. In this example, clusters 1, 2 and 3 are highlighted. The heights “H(1,2)” and “H(2,3)” represent the distances among them. In a simple analysis, as “H(1,2)” is greater than “H(2,3)”, it means that clusters 2 and 3 are closer than clusters 1 and 2.

![Figure 5 – Example of a hierarchy-agglomerative clustering](image1)

4 EXPERIMENTAL RESULTS

In this section, some results are presented regarding the vibration data obtained from experimental tests performed on a rail bridge in France. Fig. 6 shows a side view of the bridge. The studied HSR bridge is located on the South-East high speed line (kilometric point 075+317). This bridge exhibits a first natural frequency around 5 Hz which may be close to the excitation
frequency due to the train crossing (around 4-5 Hz). For this reason, this bridge was strengthened by a specific tightening system at the abutment. This tightening system allows to increase the first natural frequency in such a way that it may be shifted from the excitation one.

Three sets of dynamic tests were performed: before strengthening (15 tests), during (13 tests) and after (13 tests). The instrumentation comprises 3 vertical displacement sensors, 8 vertical accelerometers and 2 horizontal accelerometers (longitudinal and transversal) under the bridge deck, 2 temperature gauges, and 2 Q sensors which measure the axle loads at the entrance and at the clearance of the bridge. The sampling frequency was fixed at 4096Hz. The signal analysis due to train crossings highlights a good repeatability (Cremona, 2004). The idea is to make use of the clustering methods outlined in this paper in order to try to discriminate these three different stages. In other words, the goal is to separate the whole set of 41 measurements into specific groups (before, during and after). In order to perform this task, the experimental data (signals), the natural frequencies and the mode shapes of the structure will be used.

![Figure 7 – Hierarchy-divisive method applied to the signals](image1)

![Figure 8 – Dynamic clouds method applied to the signals](image2)

For the first results, Fig. 7 shows the clusters obtained by using the hierarchy-divisive method applied to the measured signals. It is possible to notice the existence of three clusters containing, globally the sets of tests related to each state of the structure. There are a few tests denoted
as “during” which were classed in wrong clusters. For example, the first cluster contains almost all the tests corresponding to the “before” state, although it also has two tests “TGV10R” and “TGV11R” which should correspond to the second cluster. It is essential to note that the tests corresponding to the “before” and “after” are identified into two distinct groups, except for the cluster “during” which presents erratic results. This observation shows that the strengthening effect is sensitive enough to be detected with this technique. Finally, it is also possible to observe the two most discriminant variables for this classification are “Sensor 10” for the first level and “Sensor 7” for the second division.

The dynamic clouds method is applied to the same data and similar results are achieved as shown in Fig. 8. It is easy to notice that all three clusters have, in its majority, one state of the structure. For this method, though, a couple of measurements are also badly classified.

For the analysis concerning the modal parameters, Table 4 shows the variation for the first 4 frequencies identified for each strengthening stages.

Table 4 - First 4 natural frequencies for each strengthening stages

<table>
<thead>
<tr>
<th>Frequencies</th>
<th>Before modification</th>
<th>During modification</th>
<th>After modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>f (Hz)</td>
<td>f (Hz)</td>
<td>f (Hz)</td>
</tr>
<tr>
<td>1</td>
<td>5.84</td>
<td>6.27</td>
<td>6.44</td>
</tr>
<tr>
<td>2</td>
<td>8.61</td>
<td>8.74</td>
<td>8.78</td>
</tr>
<tr>
<td>3</td>
<td>13.09</td>
<td>13.20</td>
<td>13.31</td>
</tr>
<tr>
<td>4</td>
<td>16.95</td>
<td>17.21</td>
<td>17.34</td>
</tr>
</tbody>
</table>

Fig. 9 shows the results obtained when a Principal Component Analysis (PCA) is applied to the symbolic data representing the frequencies. Once again, the use of symbolic data presents reasonable classification results by distinguishing different conditions of the bridge. The frequencies related to the “before” state of the bridge are grouped in the superior right region of the plot whereas the “after” state is located at the lower right part of the graph. The “during” condition shows a transition between both previous described states which is in complete agreement with the real physical circumstances.

Figure 9 – Principal Component Analysis applied to the frequencies
Furthermore, the classification achieved by using the symbolic data representing the vibration mode shapes is presented. Fig. 10 presents the results obtained when a hierarchy-agglomerative is applied to the mode shapes. The clusters are highlighted as “1”, “2” and “3” in order to show each cluster obtained. Similarly to the analysis performed with the natural frequencies, it is possible to observe a great number of mode shapes badly clustered. This can be due to the dissimilarity measure used which did not provide suitable discrimination criteria among the mode shapes.

![Image of a tree diagram with 3 clusters labeled 1, 2, and 3]

**Figure 10 – Hierarchy-agglomerative method applied to the vibration mode shapes**

5 CONCLUSIONS

In this text some principles of symbolic data analysis and, more specifically, the use of clustering methods applied to structural damage assessment have been introduced. The main idea was to use different classification procedures in order to have insights about structural health conditions. In that sense, the SDA and three clustering methods - hierarchy-divisive, dynamic clouds and hierarchy-agglomerative - were applied to either the vibration data (signals) or the natural frequencies in order to identify any structural modification processes. Experimental applications containing three structural conditions of a rail bridge were studied. The objective was to use the clustering methods to discriminate these different conditions. In order to carry out this task, both vibration data and modal parameters were used. The results obtained show that the SDA and the clustering methods considered were able to discriminate a structural modification process either considering the vibration data or the modal parameters. It was possible to observe a couple of classifications which could not achieve the expected results. In general, both the hierarchy-divisive and the dynamic clouds methods produce better results compared to those obtained by using the hierarchy-agglomerative method.

As for a future work, it is intended to be able to calculate confidence indexes for the classification obtained by these clustering methods. Therefore, it is necessary to evaluate the influence of compare new tests to the pre-defined clusters from an initial learning. These preliminary results highlight that it is essential to compare different clustering methods, since it is possible to have either good or bad clusters by using the same input data.
REFERENCES

Diday, E.; Cremona C.; Goupil F.; Afonso F.; Rahal M. 2001 – Principes d’Analyse des données symboliques et application à la détection d’anomalies sur des ouvrages publics.