Article

Unsupervised Feature Selection for Histogram-Valued Symbolic Data by Hierarchical Conceptual Clustering

Manabu Ichino [1], Kadri Umbleja [2] and Hiroyuki Yaguchi [1]

[1,*] Tokyo Denki University; ichino@dendai.ac.jp
[2] Tallinn University of Technology

Abstract: This paper presents an unsupervised feature selection method for multi-dimensional histogram-valued data. We define a multi-role measure, called the compactness, based on the concept size of given objects and/or clusters described by a fixed number of equal probability bin-rectangles. In each step of clustering, we agglomerate objects and/or clusters so as to minimize the compactness for the generated cluster. This means that the compactness plays the role of a similarity measure between objects and/or clusters to be merged. To minimize the compactness is equivalent to maximize the dis-similarity of the generated cluster, i.e., concept, against the whole concept in each step. In this sense, the compactness plays the role of cluster quality. We also show that the average compactness of each feature with respect to objects and/or clusters in several clustering steps is useful as feature effectiveness criterion. Features having small average compactness are mutually covariate, and are able to detect *geometrically thin structure* embedded in the given multi-dimensional histogram-valued data. We obtain thorough understandings of the given data by the visualization using dendrograms and scatter diagrams with respect to the selected informative features. We illustrate the effectiveness of the proposed method by using an artificial data set and real histogram-valued data sets.

Keywords: unsupervised feature selection, histogram-valued data, compactness, hierarchical conceptual clustering, multi-role measure, visualization

1. Introduction

Unsupervised feature selection is important in pattern recognition, data mining, and generally in data science (e.g., [1-4]). Solorio-Fernández et.al. [4] evaluated and discussed many filter, wrapper, and hybrid methods, and they obtained a detail taxonomy tree of unsupervised feature selection methods. They also pointed out the challenge for complex data models is one of the important theme in unsupervised feature selection. Bock and Diday [5] and Billard and Diday [6] include methods of *Symbolic Data Analysis* (SDA) for complex data models. Diday [7] presents an over view of SDA in data science, and Billard and Diday [8] present various methods to analyze symbolic data including histogram-valued data.

This paper presents an unsupervised feature selection method for mixed-type histogram-valued data by using hierarchical conceptual clustering based on the *compactness*. The compactness define the *concept size* of rectangles describing objects and/or clusters in the given feature space. In hierarchical agglomerative methods, as noted in Billard and Diday [8], we select a (dis)similarity measure between objects and we obtain a dendrogram by merging objects and/or clusters based on the selected criterion, e.g., nearest neighbor, furthest neighbor, Ward's minimum variance, or other criteria. On the other hand, the proposed method in this paper, the compactness plays not only the role of similarity measure between objects and/or clusters, but also the roles of cluster quality criterion and feature effectiveness criterion. Therefore, we can greatly simplify to realize unsupervised feature selection for complex histogram-valued symbolic data, although the proposed method may be categorized into wrapper methods.

doi:10.20944/preprints202103.0753.v1

The structure of this paper is as follows: Section 2 describes the quantile method to represent multi-dimensional distributional data. When the given p distributional features describe each of N objects, we use histogram representations for various feature types including categorical multi-value and modal multi-value types. We transform each feature value of each object to the predetermined common number m of *bins* and their *bin probabilities*. We define m+1 *quantile vectors* ordered from the *minimum quantile vector* to the *maximum quantile vector* in order to describe each object in the p dimensional histogram-valued feature space. We define m series of p dimensional *bin-rectangles* spanned by the successive quantile vectors to have common descriptions for the given objects. Then, we define the concept size of each of m bin-rectangles by the arithmetic average of p normalized bin-widths, respectively. Section 3 describes the measure of *compactness* for the merged objects and/or clusters. For an arbitrary pair of objects and for each histogram-valued feature, we define the average cumulative distribution function based on the two histogram values, and we find m+1 quantile values including the minimum and the maximum values from the obtained cumulative distribution function for each of p features. Then, we obtain m series of p-dimensional bin rectangles with predetermined bin probabilities in order to define the *Cartesian join* of the pair of objects in the p-dimensional feature space. Under the assumption of equal bin probabilities, we define a new similarity measure, the *compactness*, of a pair of objects and/or clusters as the average of m concept sizes of bin-rectangles obtained for the pair. Section 4 describes the proposed method of *hierarchical conceptual clustering* (HCC) and exploratory method of *feature selection*, and then we show the effectiveness of the proposed method by an artificial data and by three real data sets. Section 5 is a discussion for the obtained results.

2. Representation of objects by bin-rectangles

Let $\boldsymbol{U} = \{\omega_i,\ i = 1, 2,..., N\}$ be the set of given objects, and let features F_j, j = 1, 2,..., p, describe each object. Let D_j be the domain of feature F_j, j =1, 2,…, p. Then, the *feature space* is defined by

$$\boldsymbol{D}^{(p)} = D_1 \times D_2 \times \cdots \times D_p\,. \qquad (1)$$

Since we permit the simultaneous use of various feature types, we use the notation $\boldsymbol{D}^{(p)}$ for the feature space in order to distinguish it from usual p-dimensional Euclidean space $\boldsymbol{R}^p$. Each element of $\boldsymbol{D}^{(p)}$ is represented by

$$\boldsymbol{E} = E_1 \times E_2 \times \cdots \times E_p, \qquad (2)$$

where E_j, j =1, 2, … , p, is the feature value taken by the feature F_j .

2.1 Histogram-valued feature

For each object ω_i, let each feature F_j be represented by histogram value:

$$E_{ij} = \{[a_{ijk}, a_{ij(k+1)}), p_{ijk};\ k = 1, 2,..., n_{ij}\}, \qquad (3)$$

where $p_{ij1}+ p_{ij2}+\cdots+p_{ijnij}$ =1, and n_{ij} is the number of bins that compose the histogram E_{ij}. Therefore, the Cartesian product of p histogram values represents an object ω_i:

$$\boldsymbol{E}_i = E_{i1} \times E_{i2} \times \cdots \times E_{ip}. \qquad (4)$$

Since, interval-valued feature is special case of histogram feature with n_{ij} = 1 and p_{ij1} = 1, the representation of (3) is reduced to an interval:

$$E_{ij} = [a_{ij1}, a_{ij2}). \qquad (5)$$

2.2 Histogram representation of other feature types

A. Categorical multi-valued feature

Let F_j be a categorical multi-valued feature, and let E_{ij} be a value of F_j for an object ω_i. The value E_{ij} contains one or more categorical values taken from the domain D_j that is composed of finite possible categorical values. For example, E_{ij} = {“white”, “green”} is a value taken from the domain D_j = {“white”, “red”, “blue”, “green”, “black”}. For this kind feature value, we can use again a histogram. For each value in domain D_j, we assign an interval with equal width. Then, assuming uniform probability for values in multi-valued feature, we assign probabilities to each interval associated with specific value in D_j according its presence in E_{ij}. Therefore, the feature value E_{ij} = {“white”, “green”}, for example, is now represented by the histogram E_{ij} = {[0, 1)0.5, [1, 2)0, [2, 3)0, [3, 4)0.5, [4, 5)0}.

B. Modal multi-valued feature

Let $D_j = \{v_1, v_2, ..., v_n\}$ be a finite list of possible outcomes, and be the domain of a modal multi-valued feature F_j. A feature value E_{ij} for object ω_i is a subset of D_j with nonnegative measure attached to each of the values in that subset, and the sum of those nonnegative measures is *one*:

$$E_{ij} = \{ v_{ij1}, p_{ij1}; v_{ij2}, p_{ij2}; ... ; v_{ijnij}, p_{ijnij}\}, \qquad (6)$$

where$\{v_{ij1}, v_{ij2}, ..., v_{ijnij}\}\subseteq D_j$, v_{ijk} occurs with the nonnegative weight p_{ijk}, k =1, 2,…, n_{ij}, and with $p_{ij1} + p_{ij2} +...+ p_{ijnij} = 1$.

For example, E_{ij} = {“white”, 0.8; “green”, 0.2 } is a value of the modal multi-valued feature defined on the domain D_j = {“white”, “red”, “blue”, “green”, “black”}. By the same manner for the categorical multi-valued feature, we assign a same sized interval to each possible feature value from the domain D_j. The probabilities assigned to a specific feature value of the modal multi-valued feature are used as the bin probabilities of the corresponding histogram with the same bin width. Therefore, in the above example, we have a histogram representation: E_{ij} ={[0, 1)0.8, [1, 2)0, [2, 3)0, [3, 4)0.2, [4, 5)0}.

2.3 Representation of histograms by common number of quantiles

Let $\omega_i \in \boldsymbol{U}$ be the given object, and let E_{ij} in (7) be the histogram value for *a* feature F_j:

$$E_{ij} = \{[a_{ijk}, a_{ij(k+1)}), p_{ijk}; k = 1, 2,..., n_{ij}\}. \qquad (7)$$

Then, under the assumption that n_{ij} bins have uniform distributions, we define the cumulative distribution function $F_{ij}(x)$ of the histogram (7) as:

$$F_{ij}(x) = 0 \text{ for } x \le a_{ij1}$$
$$F_{ij}(x) = p_{ij1}(x - a_{ij1})/(a_{ij2} - a_{ij1}) \text{ for } a_{ij1} \le x < a_{ij2}$$
$$F_{ij}(x) = F(a_{ij1}) + p_{ij2}(x - a_{ij2})/(a_{ij3} - a_{ij2}) \text{ for } a_{ij2} \le x < a_{ij3}$$
$$\ldots\ldots$$
$$F_{ij}(x) = F(a_{nij\text{-}1}) + p_{ijnij}(x - a_{nij})/(a_{nij+1} - a_{nij}) \text{ for } a_{nij} \le x < a_{nij+1}$$
$$F_{ij}(x) = 1 \text{ for } a_{nij+1} \le x.$$

Figure 1 illustrates such a cumulative distribution function for a histogram feature value.

If we select the number m = 4 and three cut points c_1=1/4, c_2=2/4, and c_3=3/4, we can obtain three quantile values from the equations $c1 = F_{ij}(q_1)$, $c2 = F_{ij}(q_2)$, and $c3 = F_{ij}(q_3)$. Finally, we obtain four bins $[a_{ij1}, q_1)$, $[q_1, q_2)$, $[q_2, q_3)$, and $[q_3, a_{nij+1})$ and their bin probabilities $(c_1 - 0)$, $(c_2 - c_1)$, $(c_3 - c_2)$, and $(1 - c_3)$ with the same value 1/4.

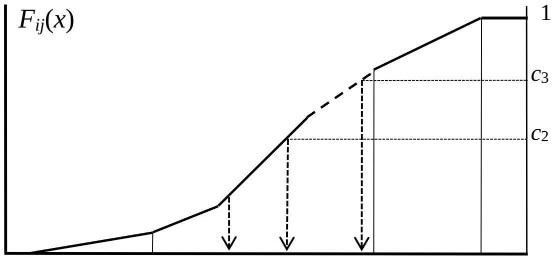


Figure 1 Cumulative distribution function and cut point probabilities.

Our general procedure to have common representation for histogram-valued data is as follows.

1) We choose common number m of quantiles.

2) Let c_1, c_2,..., $c_{m\text{-}1}$ be preselected cut points dividing the range of the distribution function $F_{ij}(x)$ into continuous intervals, i.e. bins, with preselected probabilities associated with m cut points. For example, in the quartile case we use three cut points c_1=1/4, c_2=2/4, and c_3=3/4 to have four bins with the same probability 1/4. However, we can choose different cut points, for example, c_1=1/10, c_2=5/10, and c_3=9/10 to have four bins with probabilities 1/10, 4/10, 4/10, and 1/10, respectively.

3) For the given cut points c_1, c_2,..., $c_{m\text{-}1}$, we have the corresponding quantiles by solving the following equations:

Commented [KU1]: In 1) it was c1,c2,…cm. Now it is m-1.

$$F_{ij}(x_{ij0}) = 0, \text{ (i.e. } x_{ij0} = a_{ij1})$$
$$F_{ij}(x_{ij1}) = c_1,\ F_{ij}(x_{ij2}) = c_2, \ldots,\ F_{ij}(x_{ij(m\text{-}1)}) = c_{m\text{-}1}, \text{ and}$$
$$F_{ij}(x_{ijm}) = 1, \text{ (i.e. } x_{ijm} = a_{ijnij+1}).$$

Therefore, we describe each object $\omega_i \in \boldsymbol{U}$ for each feature F_j by a (m+1) tuple:

$$(x_{ij0}, x_{ij1}, x_{ij2}, \ldots, x_{ij(m\text{-}1)}, x_{ijm}), \quad j = 1, 2, \ldots, p, \tag{8}$$

and the corresponding histogram by:

$$E_{ij} = \{[x_{ijk}, x_{ij(k+1)}), (c_{k+1}\text{-}c_k); k = 0, 1, \ldots, m\text{-}1\}, j = 1, 2, \ldots, p, \tag{9}$$

where we assume that $c_0 = 0$ and $c_m = 1$. In (9), (c_{k+1} - c_k), k = 0, 1,..., m-1, denote bin probabilities by the preselected cut point probabilities c_1, c_2,..., $c_{m\text{-}1}$. In the quartile case again, m = 4 and c_1=1/4, c_2=2/4, and c_3=3/4, and four bins [x_{ij0}, x_{ij1}), [x_{ij1}, x_{ij2}), [x_{ij2}, x_{ij3}), and [x_{ij3}, x_{ij4}) have the same probability 1/4.

It should be noted that the number of bins of the given histograms are mutually different in general. However, we can obtain (m+1)-tuples as the common representation for all histograms by selecting an integer m and a set of cut points.

2.4 Quantile vectors and bin-rectangles

For each object $\omega_i \in \boldsymbol{U}$, we define ($m$+1) p-dimensional numerical vectors, called the *quantile vectors*, as follows.

$$\boldsymbol{x}_{ik} = (x_{i1k}, x_{i2k}, \ldots, x_{ipk}),\ k = 0, 1, \ldots, m. \tag{10}$$

We call $\boldsymbol{x}_{i0}$ and $\boldsymbol{x}_{im}$ the minimum quantile vector and the maximum quantile vector for $\omega_i \in \boldsymbol{U}$, respectively. Therefore, m+1 quantile vectors $\{\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im}\}$ in $\boldsymbol{R}^p$ describe each object $\omega_i \in \boldsymbol{U}$ together with cut point probabilities.

The components of m+1 quantile vectors in (10) for object $\omega_i \in \boldsymbol{U}$ satisfy the inequalities:

$$x_{ij0} \le x_{ij1} \le x_{ij2} \le \cdots \le x_{ij(m\text{-}1)} \le x_{ijm},\ j = 1, 2, \ldots, p. \tag{11}$$

Therefore, m+1 quantile vectors in (10) for object $\omega_i \in \boldsymbol{U}$ satisfy the monotone property:

$$\boldsymbol{x}_{i0} \le \boldsymbol{x}_{i1} \le \cdots \le \boldsymbol{x}_{im}. \quad (12)$$

For the series of quantile vectors $\boldsymbol{x}_{i0}$, $\boldsymbol{x}_{i1}$,..., $\boldsymbol{x}_{im}$ of object $\omega_i \in \boldsymbol{U}$, we define m series of p dimensional rectangles spanned by adjacent quantile vectors $\boldsymbol{x}_{ik}$ and $\boldsymbol{x}_{i(k+1)}$, k = 0, 1,..., m-1, as follows:

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{x}_{ik}, \boldsymbol{x}_{i(k+1)}) &= \boldsymbol{x}_{ik} \oplus \boldsymbol{x}_{i(k+1)} = (x_{i1k} \oplus x_{i1(k+1)}) \times (x_{i2k} \oplus x_{i2(k+1)}) \times \cdots \times (x_{ipk} \oplus x_{ip(k+1)})\\ &= [x_{i1k}, x_{i1(k+1)}] \times [x_{i2k}, x_{i2(k+1)}] \times \cdots \times [x_{ipk}, x_{ip(k+1)}],\ k = 0, 1,..., m\text{-}1, \quad (13)\end{aligned}$$

where $\boldsymbol{x}_{ik} \oplus \boldsymbol{x}_{i(k+1)}$ is the Cartesian join (Ichino and Yaguchi [13]) of $\boldsymbol{x}_{ik}$ and $\boldsymbol{x}_{i(k+1)}$ obtained by the Cartesian join $x_{ijk} \oplus x_{ij(k+1)} = [x_{ijk}, x_{ij(k+1)}]$, j = 1, 2,..., p, and we call $\boldsymbol{B}(\boldsymbol{x}_{ik}, \boldsymbol{x}_{i(k+1)})$, k = 0, 1,..., m-1, as the *bin-rectangles*.

Figure 2 illustrates two objects ω_i and ω_l by the representations by bin-rectangles of quartile case in two-dimensional Euclidean space. Since a p-dimensional rectangle in $\boldsymbol{R}^p$ is equivalent to a conjunctive logical expression, e.g., Michalski and Step [9], we use also the term *concept* for a rectangular expression in the space $\boldsymbol{R}^p$. In other word, m bin-rectangles describe each of objects ω_i and ω_l as concepts. We should note that the selection of a larger value m yields smaller rectangles as possible descriptions. In this sense, the selection of the integer number m controls the granularity of concept descriptions.

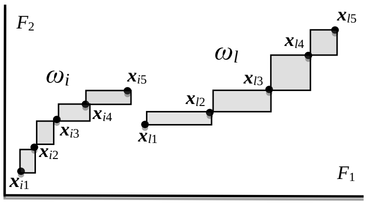


Figure 2 Representations of objects by bin-rectangles in the quartile case.

2.5 Concept size of bin-rectangles

For each feature F_j, j = 1, 2,…, p, let the domain D_j of feature values be the following interval:

$D_j = [x_{jmin}, x_{jmax}]$, j = 1, 2, …, p, where

$x_{jmin} = min(x_{1j0}, x_{2j0},…, x_{Nj0})$ and $x_{jmax} = max(x_{1jm}, x_{2jm},…, x_{Njm})$.

Definition 1

Let an object $\omega_i \in \boldsymbol{U}$ be described by the set of histograms E_{ij} in (9). We define the average concept size $P(E_{ij})$ of m bins for histogram E_{ij} by

$$\begin{aligned}P(E_{ij}) &= \{c_1(x_{ij1} - x_{ij0}) + (c_2 - c_1)(x_{ij2} - x_{ij1}) + \cdots + (c_k + c_{(k-1)})(x_{ijk} - x_{ij(k-1)}) + \cdots\\ &\quad + (c_{m-1} - c_{m-2})(x_{ij(m-1)} - x_{ij(m-2)}) + (1 - c_{m-1})(x_{ijm} - x_{ij(m-1)})\}/|D_j|,\\ &= \{c_1|x_{ij0} \oplus x_{ij1}| + (c_2 - c_1)|x_{ij1} \oplus x_{ij2}| + \cdots + (c_k + c_{(k-1)})|x_{ij(k-1)} \oplus x_{ijk}| + \cdots\\ &\quad + (c_{m-1} - c_{m-2})|x_{ij(m-1)} \oplus x_{ij(m-2)}| + (1 - c_{m-1})|x_{ijm} \oplus x_{ij(m-1)}|\}/|D_j|,\ j = 1, 2, ..., p, \quad (14)\end{aligned}$$

where $x_{ij(k-1)} \oplus x_{ijk}$ defines the Cartesian join of $x_{ij(k-1)}$ and x_{ijk} as the interval spanned by them, and where $|D_j|$ and $|x_{ij(k-1)} \oplus x_{ijk}|$ are the length of the domain and the k-th bin, respectively

The average concept size $P(E_{ij})$ satisfies the inequality:

$$0 \le P(E_{ij}) \le 1,\ j = 1, 2,..., p. \quad (15)$$

Example 1

1) When E_{ij} is a histogram with a single bin, the concept size is

$P(E_{ij}) = (x_{ij1} - x_{ij0})/|D_j|$.

2) When E_{ij} is a histogram with *four* bins with equal probabilities, i.e. a quartile case, the average concept size of four bins is $P(E_{ij}) = (x_{ij4} - x_{ij0})/(4|D_j|)$.

3) When E_{ij} is a histogram with four bins with cut points $c_1 = 1/10$, $c_2 = 5/10$, and $c_3 = 9/10$, the average concept size of *four* bins is

$$P(E_{ij}) = \{(x_{ij1} - x_{ij0})/10 + 4(x_{ij2} - x_{ij1})/10 + 4(x_{ij3} - x_{ij2})/10 + (x_{ij4} - x_{ij3})/10\}/|D_j|$$
$$= (x_{ij4} + 3x_{ij3} - 3x_{ij1} - x_{ij0})/(10|D_j|).$$

4) In the Hardwood data (see Section 4.4), *seven* quantile values for *five* cut point probabilities, c_1=1/10, c_2=1/4, c_3=1/2, c_4=3/4, and c_5=9/10, describe each histogram E_{ij}. Then the average concept size of *six* bins becomes:

$$P(E_{ij}) = \{(10(x_{ij1} - x_{ij0})/100 + 15(x_{ij2} - x_{ij1})\ /100 + 25(x_{ij3} - x_{ij2})/100 + 25(x_{ij4} - x_{ij3})/100$$
$$+ 15(x_{ij5} - x_{ij4})/100 + 10(x_{ij6} - x_{ij5})/100\}/|D_j|$$
$$= \{10x_{ij6} + 5x_{ij5} + 10x_{ij4} - 10x_{ij2} - 5x_{ij1} -10x_{ij0}\}/(100|D_j|)$$
$$= \{2x_{ij6} + x_{ij5} + 2x_{ij4} - 2x_{ij2} - x_{ij1} -2x_{ij0}\}/(20|D_j|)$$

This example asserts the simplicity of concept size in case of equal bin probabilities.

Proposition 1

1) When m bin probabilities are the same, the average concept size of m bins is reduced to the form:

$$P(E_{ij}) = (x_{ijm} - x_{ij0})\ /(m|D_j|), \quad j = 1, 2,..., p \qquad (16)$$

2) When m bin widths are the same size w_{ij}, we have:

$$P(E_{ij}) = w_{ij}/|D_j|,\ j = 1, 2,..., p, \qquad (17)$$

3) It is clear that:

$$w_{ij} = (x_{ijm} - x_{ij0})/m. \qquad (18)$$

Proof of Proposition 1. Since m bin probabilities are the same, we have

$$c_1 = (c_2 - c_1) = \cdots = (c_{m-1} - c_{m-2}) = (1 - c_{m-1}) =1/m.$$

Then, (14) leads to (16). On the other hand, m bin widths are the same size w_{ij}, we have

$$c_1w_{ij} + (c_2 - c_1)w_{ij} + \cdots + (c_{m-1} - c_{m-2})w_{ij} + (1 - c_{m-1})w_{ij} = w_{ij}.$$

Then, (14) leads to (17). (18) is clear, since mw_{ij} equals to the span $(x_{ijm} - x_{ij0})$. □

This proposition asserts that the both extremes yield the same conclusion.

Definition 2

Let $\boldsymbol{E}_i = E_{i1} \times E_{i2} \times \cdots \times E_{ip}$ be the description by p histograms in $\boldsymbol{R}^p$ for $\omega_i \in \boldsymbol{U}$.

Then, we define the concept size $P(\boldsymbol{E}_i)$ of $\boldsymbol{E}_i$ by the arithmetic mean

$$P(\boldsymbol{E}_i) = (P(E_{i1}) + P(E_{i2}) + \cdots + P(E_{ip}))/p. \qquad (19)$$

From (15), It is clear that:

$$0 \le P(\boldsymbol{E}_i) \le 1. \qquad (20)$$

Definition 3

Let $P(\boldsymbol{B}(\boldsymbol{x}_{ik}, \boldsymbol{x}_{i(k+1)}))$, k = 0, 1,..., m-1, be the concept size of m bin-rectangles defined by the average of p normalized bin-widths:

$$P(\boldsymbol{B}(\boldsymbol{x}_{ik}, \boldsymbol{x}_{i(k+1)}))=\{|x_{i1k} \oplus x_{i1(k+1)}|/|D_1|+|x_{i2k} \oplus x_{i2(k+1)}|/|D_2|+\cdots+|x_{ipk} \oplus x_{ip(k+1)}|/|D_p|\}/p,\ k=0,1,...,\ m-1. \qquad (21)$$

Then (14) and (21) lead to the following proposition.

Proposition 2

The concept size $P(\boldsymbol{E}_i)$ is equivalent to the average value of m concept sizes of bin-rectangles:

$$P(\boldsymbol{E}_i) = (c_1 - c_0)P(\boldsymbol{B}(\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1})) + (c_2 - c_1)P(\boldsymbol{B}(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})) + \cdots + (c_m - c_{(m-1)})P(\boldsymbol{B}(\boldsymbol{x}_{i(m-1)}, \boldsymbol{x}_{im})), \qquad (22)$$

where $c_0 = 0$ and $c_m = 1$.

In Figure 2, two objects ω_i and ω_l are represented by four bin-rectangles with the same probability 1/4. Hence, smaller sized bin-rectangles mean that they have higher probability densities with respect to features under consideration. In this sense, object ω_i has a sharp probability distribution compared that of object ω_l. By the virtue of equiprobability assumption, we can easily compare the object descriptions by series of bin-rectangles under the selected feature sub-space. If we use the descriptions of objects under the assumption of equal bin widths, we can no longer compare between objects in such a simple way.

3. Concept size of the Cartesian join of objects and the compactness

3.1 Concept size of the Cartesian join of objects

A major merit of the quantile representation is that we are able to have a common numerical representation for various types of histogram data. We select a common integer number m, then we obtain common form of histograms with m bins and the predetermined bin probabilities for each of p features describing each object.

Let E_{ij} and E_{lj} be two histogram values of objects ω_i , $\omega_l \in \boldsymbol{U}$ with respect to the j-th feature. We represent a generalized histogram value of E_{ij} and E_{lj}, called the Cartesian join of E_{ij} and E_{lj}, by $E_{ij} \oplus E_{lj}$. Let $F_{Eij}(x)$ and $F_{Elj}(x)$ be the cumulative distribution functions associated with histograms E_{ij} and E_{lj}, respectively.

Definition 4

We define the cumulative distribution function for the Cartesian join $E_{ij} \oplus E_{lj}$ by

$$F_{Eij \oplus Elj}(x) = (F_{Eij}(x) + F_{Elj}(x))/2,\ j = 1, 2,..., p. \qquad (23)$$

Then, by applying the same integer number m and the set of cut point probabilities, $c_1, c_2,..., c_{m-1}$, used for E_{ij} and E_{lj}, we define the histogram of the Cartesian join $E_{ij} \oplus E_{lj}$ for j-th feature as:

$$E_{ij} \oplus E_{lj} = \{[x_{(i+l)jk}, x_{(i+l)j(k+1)}), (c_{k+1} - c_k);\ k = 0, 1,..., m-1\},\ j = 1, 2,..., p, \qquad (24)$$

where we assume that $c_0 = 0$ and $c_m = 1$ and that the suffix $(i+l)$ denotes the quantile values for the Cartesian join $E_{ij} \oplus E_{lj}$. We should note that $x_{(i+l)j0} = min(x_{ij0}, x_{lj0})$ and $x_{(i+l)jm} = max(x_{ijm}, x_{ljm})$.

Definition 5

We define the average concept size $P(E_{ij} \oplus E_{lj})$ of m bins for the Cartesian join E_{ij} and E_{lj} under the j-th feature as follows.

$$\begin{aligned} P(E_{ij} \oplus E_{lj}) = \{&c_1(x_{(i+l)j1} - x_{(i+l)j0}) + (c_2 - c_1)(x_{(i+l)j2} - x_{(i+l)j1}) + \cdots \\ &+ (c_{m-1} - c_{m-2})(x_{(i+l)j(m-1)} - x_{(i+l)j(m-2)}) + (1 - c_{m-1})(x_{(i+l)jm} - x_{(i+l)j(m-1)})\}/|D_j|, \\ = \{ &c_1|x_{(i+l)j0} \oplus x_{(i+l)j1}| + (c_2 - c_1)|x_{(i+l)j1} \oplus x_{(i+l)j2}| + \cdots \\ &+ (c_{m-1} - c_{m-2})|x_{(i+l)j(m-2)} \oplus x_{(i+l)j(m-1)}| + (1 - c_{m-1})|x_{(i+l)j(m-1)} \oplus x_{(i+l)jm}|\}/ |D_j|,\ j = 1, 2,..., p. \qquad (25) \end{aligned}$$

The average concept size $P(E_{ij} \oplus E_{lj})$ satisfies the inequality:

$$0 \leq P(E_{ij} \oplus E_{lj}) \leq 1,\ j = 1, 2,..., p. \qquad (26)$$

Proposition 3

When m bin probabilities are the same or m bin widths are the same, we have the following monotone property.

$$P(E_{ij}), P(E_{lj}) \leq P(E_{ij} \oplus E_{lj}),\ j = 1, 2,..., p. \qquad (27)$$

Proof of Proposition 3. If the bin probabilities are the same with the value $1/m$, (25) becomes simply as

$$P(E_{ij}\oplus E_{lj}) = (x_{(i+l)jm} - x_{(i+l)j0})/(m|D_j|),\ j = 1, 2, ..., p.$$

Then, the following inequality leads to the result (27).

$$(x_{ijm} - x_{ij0})/(m|D_j|),\ (x_{ljm} - x_{lj0})/(m|D_j|) \le (max(x_{ijm}, x_{ljm}) - min(x_{ij0}, x_{lj0}))/(m|D_j|). \tag{28}$$

On the other hand, from Proposition 1, (28) is equivalent to $w_{ij}/|D_j|$, $w_{lj}/|D_j| \le w_{(i+l)j}/|D_j|$. Hence, we have (27). □

Definition 6

Let $\boldsymbol{E}_i = E_{i1} \times E_{i2} \times \cdots \times E_{ip}$ and $\boldsymbol{E}_l = E_{l1} \times E_{l2} \times \cdots \times E_{lp}$ be the descriptions by p histograms in $\boldsymbol{R}^p$ for ω_i and ω_l, respectively. Then, we define the concept size $P(\boldsymbol{E}_i\oplus\boldsymbol{E}_l)$ for the Cartesian join of $\boldsymbol{E}_i$ and $\boldsymbol{E}_l$ by the arithmetic mean

$$P(\boldsymbol{E}_i\oplus\boldsymbol{E}_l) = (P(E_{i1}\oplus E_{l1}) + P(E_{i2}\oplus E_{l2}) + \cdots + P(E_{ip}\oplus E_{lp}))/p. \tag{29}$$

From (26), It is clear that:

$$0 \le P(\boldsymbol{E}_i\oplus\boldsymbol{E}_l) \le 1. \tag{30}$$

Definition 7

Let $\boldsymbol{x}_{(i+l)k}$, $k = 0, 1,..., m$, be the quantile vectors for the Cartesian join $\boldsymbol{E}_i\oplus\boldsymbol{E}_l$, and let $P(\boldsymbol{B}(\boldsymbol{x}_{(i+l)k}, \boldsymbol{x}_{(i+l)(k+1)}))$, $k = 0, 1,..., m\text{-}1$, be the concept sizes of m bin-rectangles defined by the average of p normalized bin-widths:

$$P(\boldsymbol{B}(\boldsymbol{x}_{ik}, \boldsymbol{x}_{i(k+1)}))=\{|x_{i1k}\oplus x_{i1(k+1)}|/|D_1|+|x_{i2k}\oplus x_{i2(k+1)}|/|D_2|+\cdots+|x_{ipk}\oplus x_{ip(k+1)}|/|D_p|\}/p,\ k = 0,1,..., m\text{-}1. \tag{31}$$

Then we have the following result by the same way in Proposition 2.

Proposition 4

The concept size $P(\boldsymbol{E}_i\oplus\boldsymbol{E}_l)$ is equivalent to the average value of m concept sizes of bin-rectangles:

$$P(\boldsymbol{E}_i\oplus\boldsymbol{E}_l) = (c_1 - c_0)P(\boldsymbol{B}(\boldsymbol{x}_{(i+l)0}, \boldsymbol{x}_{(i+l)1})) + (c_2 - c_1)P(\boldsymbol{B}(\boldsymbol{x}_{(i+l)1}, \boldsymbol{x}_{(i+l)2})) + \cdots$$
$$+ (c_m\text{-}c_{(m-1)})P(\boldsymbol{B}(\boldsymbol{x}_{(i+l)(m-1)}, \boldsymbol{x}_{(i+l)m})), \tag{32}$$

where $c_0 = 0$ and $c_m =1$.

We have the following monotone property from Proposition 3 and Definition 6.

Proposition 5

When m bin probabilities are the same or m bin widths are the same for all features, we have the monotone property:

Commented [KU2]: Very importan!

$$P(\boldsymbol{E}_i),\ P(\boldsymbol{E}_l) \le P(\boldsymbol{E}_i\oplus\boldsymbol{E}_l). \tag{33}$$

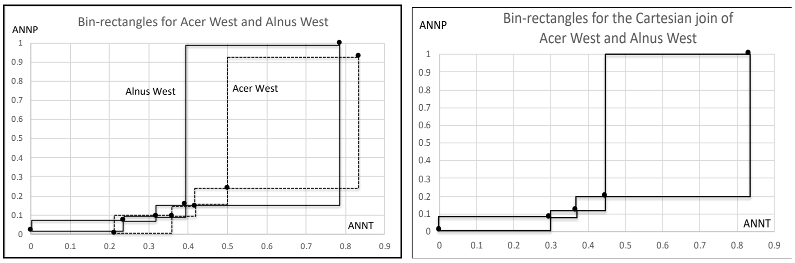
This property plays a very important role in our hierarchical conceptual clustering.

Example 2

Table 1 shows two hardwoods, Acer West and Alnus West, under quartile descriptions for two features, Anual Temerature (ANNT) and Anual Precipitation (ANNP), by *zero-one* normalized feature values under the selected *ten* harwoods used in Section 4.4. Figure 3 (a) shows the descriptions of Acer West and Alnus West by four series of bin-rectangles. The fourth bin-rectangles for both hardwoods are very large. Hence, they have very low probability density compared to other bin-rectangles. Figure 3 (b) is the description by bin-rectangles for the Cartesian join of Acer West and Alnus west.

Table 1 Two hardwoods by quartile representations.

Quantiles	ANNT	ANNP
Acer West 0	0.211	0.004
1	0.358	0.091
2	0.416	0.145
3	0.500	0.237
4	0.832	0.932
Alnus West 0	0.000	0.018
1	0.234	0.071
2	0.317	0.092
3	0.391	0.153
4	0.784	1.000



(a) (b)

Figure 3 Bin-rectangles for two hardwoods (a) and their Cartesian join (b).

Table 2 The average concept sizes for two hardwoods and their Cartesian join.

Concept size	ANNT	ANNP	Average
Acer West	0.155	0.232	0.194
Alnus West	0.196	0.245	0.221
Cartesian join	0.208	0.249	0.229

Table 2 shows the average concept sizes for each hard wood and for each feature by Definition 1, and shows also the average concept sizes by Definitions 2 and 3. We can confirm the monotone properties in Propositions 3 and 5. The Cartesian join of two hardwoods for ANNP achieves almost the maximum concept size 1/4. Therefore, four bin-intervals of the Cartesian join for ANNP almost span the whole interval [0, 1].

3.2 *Compactness* and its properties

In the following, we assume that the given distributional data having the same representation by *m* quantile values with the same bin probabilities, since we can confirm the monotone property in Propositions 3 and 5, and we can easily visualize objects and their Cartesian joins by bin-rectangles under the selected features as in Figure 2.

Definition 8

Under the assumption of equal bin probabilities, we define the compactness of the generalized concept by ω_i and ω_l as:

$$C(\omega_i, \omega_l) = P(\boldsymbol{E}_i \oplus \boldsymbol{E}_l) = (P(\boldsymbol{B}(\mathbf{x}_{(i+l)0}, \mathbf{x}_{(i+l)1}))+P(\boldsymbol{B}(\mathbf{x}_{(i+l)1}, \mathbf{x}_{(i+l)2}))+\cdots+P(\boldsymbol{B}(\mathbf{x}_{(i+l)(m-1)}, \mathbf{x}_{(i+l)m}))/m. \quad (34)$$

In Figure 3 (b), the Cartesian join of Acer West and Alnus West is the series of four bin-rectangles and the compactness of Acer West and Alnus West is the average value of the concept sizes of four bin-rectangles in (b). Therefore, the concept size of fourth bin determine almost the concept size.

The compactness satisfies the following properties.

Proposition 6

1) $0 \leq C(\omega_i, \omega_l) \leq 1$

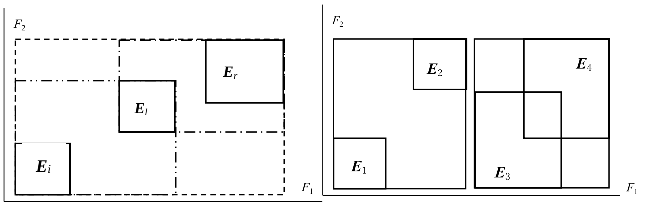
2) $C(\omega_i, \omega_l) = 0$ *iff* $\boldsymbol{E}_i \equiv \boldsymbol{E}_l$ and has null size ($P(\boldsymbol{E}_i) = 0$)

3) $C(\omega_i, \omega_i), C(\omega_l, \omega_l) \leq C(\omega_i, \omega_l)$

4) $C(\omega_i, \omega_l) = C(\omega_l, \omega_i)$

5) $C(\omega_i, \omega_r) \leq C(\omega_i, \omega_l) + C(\omega_l, \omega_r)$ may not hold in general.

Proof of Proposition 6. Definitions 6 and 7, and Propositions 4 and 5 lead 1) ~ 4). Figure 4 (a) is a counter example for 5). □



(a) (b)

Figure 4 Examples for compactness.

Figure 4 (b) illustrates the Cartesian join for interval valued objects. We should note that the compactness $C(\omega_1, \omega_2) = P(\boldsymbol{E}_1 \oplus \boldsymbol{E}_2)$ and $C(\omega_3, \omega_4) = P(\boldsymbol{E}_3 \oplus \boldsymbol{E}_4)$ take the same value as the concept size. On the other hand, we usually expect that any (dis)similarity measures for distributional data should take different values for the pairs ($\boldsymbol{E}_1$, $\boldsymbol{E}_2$) and ($\boldsymbol{E}_3$, $\boldsymbol{E}_4$). Therefore, a small value compactness requires that the pair of objects under consideration should be similar each other, but the converse is not true.

In hierarchical conceptual clustering, the compactness is useful as the measure of similarity between objects and/or clusters. We merge objects and/or clusters so as to minimize the compactness. This means also to maximize the dissimilarity against the whole concept. Therefore, the compactness plays dual roles as a similarity measure and a measure of cluster quality.

4. Exploratory hierarchical concept analysis

This section describes our algorithm of hierarchical conceptual clustering and an exploratory method for unsupervised feature selection based on the concept size and our clustering method. We use an artificial data and Oil's data to explain our exploratory hierarchical method. We also analyze two other real data sets in order to show the usefulness of the proposed method.

4-1. Hierarchical conceptual clustering

Let $\boldsymbol{U} = \{\omega_1, \omega_2, \ldots, \omega_N\}$ be the given set of objects, and let each object ω_i be described by a set of histograms $\boldsymbol{E}_i = E_{i1} \times E_{i2} \times \cdots \times E_{ip}$ in the feature space $\boldsymbol{R}^p$. We assume that all histogram values for all objects have the same number m of quantiles and the same bin probabilities.

Algorithm (Hierarchical Conceptual Clustering (HCC)

Step 1: For each pair of objects ω_i and ω_l in $\boldsymbol{U}$, evaluate the compactness $C(\omega_i, \omega_l)$ and find the pair ω_q and ω_r that minimizes the compactness.

Step 2: Add the merged concept ω_{qr} ={ω_q, ω_r} to ***U*** and delete ω_q and ω_r from ***U***, where the representation of ω_{qr} follows to the Cartesian join in Definition 4 under the assumption of *m* quantiles and the equal bin probabilities.

Step 3: Repeat Step 1 and Step 2 until ***U*** includes only one concept, i.e., the whole concept.

4-2. An exploratory method of feature selection

We use an artificial data and Oils' data (Ichino and Umbleja[18]) to illustrate feature selection capability to extract a covariate feature subset in which the given data sets take "geometrically thin structures" (Ono and Ichino[14]).

1) Artificial data

Sixteen small rectangles organize an oval structure in the first two features F1 and F2 as shown in Figure 5. For each of sixteen objects, we transform the feature values of F1 and F2 to 0-1 normalized interval values. Then, we select additional three randomly selected interval values in the unit interval [0, 1] for features F3, F4, and F5. Table 3 summarizes sixteen objects described by five 0-1 normalized interval valued features. It should be noted that usual numerical data is regarded as special type of interval data, i.e., null interval.

Figure 6 shows the result by the quantile method of PCA (Ichino[15]). Each numbered arrow line connects the minimum and the maximum quantile vectors, and describes the corresponding rectangular object. The oval structure embedded in the first two features cannot be reproduced in the factor plane. Any well-known correlation criterion fails to capture the embedded oval structure.

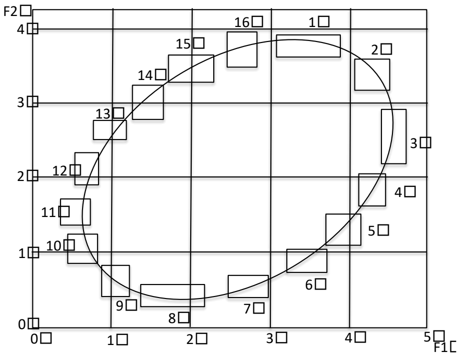


Figure 5 Oval data.

Table 3 Oval artificial data.

	F1	F2	F3	F4	F5	Concept size
1	[0.629, 0.798]	[0.905, 0.986]	[0.000, 0.982]	[0.002, 0.883]	[0.360, 0.380]	0.427
2	[0.854, 0.955]	[0.797, 0.905]	[0.002, 0.421]	[0.573, 1.000]	[0.754, 0.761]	0.212
3	[0.921, 1.000]	[0.527, 0.716]	[0.193, 0.934]	[0.035, 0.477]	[0.406, 0.587]	0.326
4	[0.865, 0.933]	[0.378, 0.500]	[0.452, 0.854]	[0.213, 0.604]	[0.000, 0.074]	0.211
5	[0.775, 0.876]	[0.257, 0.338]	[0.300, 0.614]	[0.425, 0.979]	[0.217, 0.568]	0.280
6	[0.663, 0.764]	[0.135, 0.216]	[0.712, 1.000]	[0.904, 0.968]	[0.103, 0.950]	0.276
7	[0.494, 0.596]	[0.041, 0.122]	[0.293, 0.470]	[0.023, 0.086]	[0.765, 0.902]	0.112
8	[0.225, 0.427]	[0.000, 0.081]	[0.633, 0.872]	[0.000, 0.582]	[0.719, 0.852]	0.247
9	[0.112, 0.213]	[0.041, 0.149]	[0.167, 0.802]	[0.056, 0.129]	[0.124, 0.642]	0.287
10	[0.022, 0.112]	[0.162, 0.270]	[0.026, 0.718]	[0.418, 0.851]	[0.549, 0.853]	0.325
11	[0.000, 0.090]	[0.297, 0.392]	[0.096, 0.759]	[0.438, 0.938]	[0.495, 0.760]	0.323
12	[0.045 ,0.112]	[0.446, 0.554]	[0.826, 0.962]	[0.230, 0.755]	[0.104, 0.189]	0.184
13	[0.101, 0.202]	[0.608, 0.676]	[0.367, 0.570]	[0.236, 0.684]	[0.683, 0.930]	0.213
14	[0.213, 0.292]	[0.676, 0.811]	[0.371, 0.381]	[0.086, 0.305]	[0.009, 1.000]	0.287
15	[0.315, 0.438]	[0.811, 0.919]	[0.049, 0.585]	[0.056, 0.891]	[0.528, 0.881]	0.391
16	[0.483, 0.562]	[0.878, 1.000]	[0.402, 0.609]	[0.150, 0.769]	[0.207, 0.732]	0.310
Average CS	0.103	0.105	0.415	0.441	0.315	0.276

Figure 7 is the dendrogram based on the compactness for the first two features. It is clear that each cluster grows up along the oval structure of Figure 5. Our HCC generate *eight* comparable sized rectangles along the oval structure, then generate *four* rectangles, and so on. On the other hand, Figure 8 is the dendrogram for the given five features. We can also recognize the fact that each cluster grows up again along the oval structure of Figure 5 in spite of the addition of three useless features.

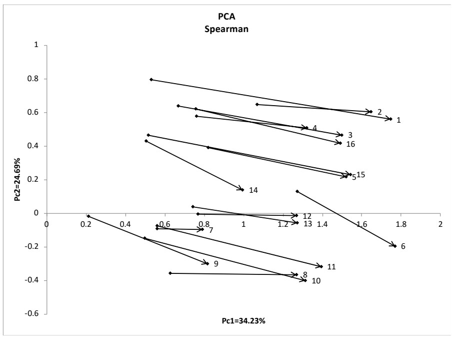


Figure 6 PCA result for Oval artificial data.

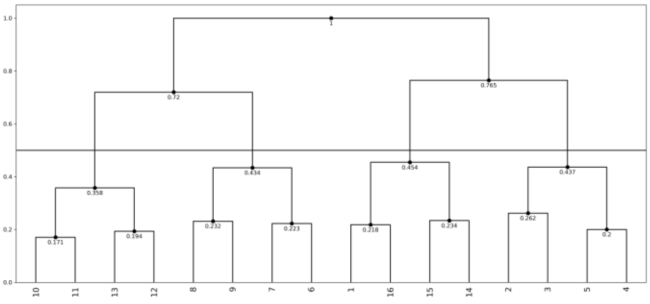


Figure 7 Dendrogram by the HCC for the first two features.

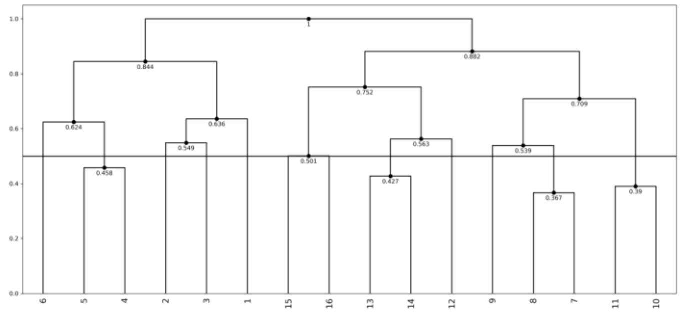


Figure 8 Dendrogram by the HCC for five features.

Table 4 summarizes the average compactness for each feature in each step of hierarchical clustering. For example, in Step 1, our HCC generates a larger rectangle by objects 10 and 11. Then, for each feature, we recalculate the average side lengths of 15 rectangles including enlarged rectangle. The result is the second row in Table 4. We repeat the same procedure for succeeding clustering steps. Until step 13, i.e., number of clusters are three, the importance of the first two features are valid. In many steps, the values of average compactness of features F1 and F2 are sufficiently small compared to the middle point value 0.5. On the other hand, the values of average compactness for other three features grows up rapidly exceeding the middle point 0.5. Thus, we conclude that the first two features are robustly informative through the clustering process. The proposed method could detect our oval structure embedded in five dimensional interval valued data as a *geometrically thin structure*. In the following, we use the middle point 0.5 as a criterion whether the average compactness is small or large.

Commented [KU3]: I think this is very good and valuable point here!

Table 4 Average compactness of each feature in each clustering step.

Clustering step	Average Compactness				
	F1	F2	F3	F4	F5
0	**0.103**	**0.105**	0.415	0.441	0.315
1	**0.115**	**0.109**	0.454	0.466	0.330
2	**0.118**	**0.119**	0.442	0.470	0.338
3	**0.128**	**0.128**	0.475	0.501	0.345
4	**0.138**	**0.142**	0.501	0.528	0.386
5	**0.154**	**0.151**	0.530	0.519	0.403
6	**0.171**	**0.158**	0.532	0.564	0.451
7	**0.186**	**0.185**	0.566	0.637	0.519
8	**0.208**	**0.215**	0.669	0.660	0.574
9	**0.239**	**0.251**	0.744	0.744	0.589
10	**0.288**	**0.293**	0.712	0.727	0.692
11	**0.346**	**0.354**	0.736	0.839	0.759
12	**0.438**	**0.443**	0.860	0.882	0.780
13	**0.494**	**0.599**	0.919	0.924	0.906
14	**0.483**	0.926	0.967	0.968	0.971
15	1.000	1.000	1.000	1.000	1.000

2) Oils' data

The data in Table 5 describes six plant oils; Linseed, Perilla, Cotton, Sesame, Camellia, and Olive, and two fats; Beef and Hog, by five interval valued features; Specific gravity, Freezing point, Iodine value, Saponification value, and Major acids.

Table 5 Oils' data [18].

	Specific gravity	Freezing point	Iodine value	Saponification v.	Major acids
Linseed	[0.930, 0.935]	[-27, -18]	[170, 204]	[118, 196]	[1.75, 4.81]
Prilla	[0.930, 0.937]	[-5, -4]	[192, 208]	[188, 197]	[0.77, 4.85]
Cotton	[0.916, 0.918]	[-6, -1]	[99, 113]	[189, 198]	[0.42, 3.84]
Sesame	[0.920, 0.926]	[-6,-4]	[104, 116]	[187, 193]	[0.91, 3.77]
Camellia	[0.916, 0.917]	[-21, -15]	[80, 82]	[189, 193]	[2.00, 2.98]
Olive	[0.914, 0.919]	[0, 6]	[79, 90]	[187, 196]	[0.83, 4.02]
Beef	[0.860, 0.870]	[30, 38]	[40, 48]	[190, 199]	[0.31, 2.89]
Hog	[0.858, 0.864]	[22, 32]	53, 77]	[190, 202]	[0.37, 3.65]

The result of PCA in Figure 9 and the dendrogram in Figure 10 show three explicit clusters (Linseed, Perilla), (Cotton, Sesame, Olive, Camellia), and (Beef, Hog). Table 6 summarizes the values of the average compactness for each feature in each clustering step. We see that the most robustly informative feature is *Specific gravity* and then *Iodine value* until Step 5 obtaining three clusters. In initial step, *major acids* exceeds our basic criterion 0.5.

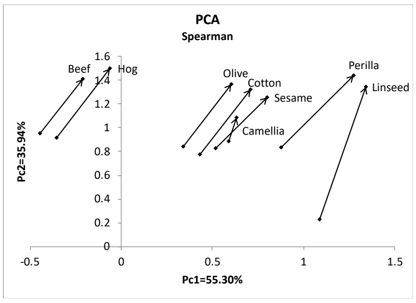


Figure 9　PCA result for Oil's data.

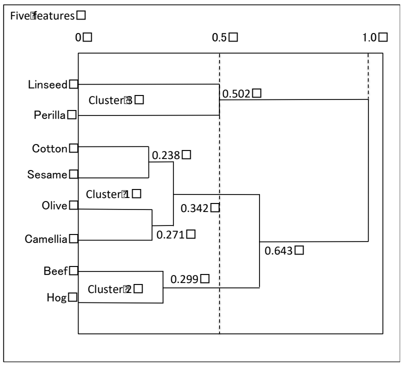


Figure 10　Dendrogram by HCC for five features.

Table 6　Average compactness of each feature in each clustering step.

Feature	Average compactness for each clustering step							
	0	1	2	3	4	5	6	7
Specific gravity	0.066	0.080	0.091	0.099	0.114	0.131	0.475	1
Freezing point	0.090	0.099	0.154	0.178	0.204	0.338	0.631	1
Iodine value	0.090	0.095	0.109	0.137	0.185	0.222	0.339	1
Saponification value	0.202	0.224	0.254	0.283	0.327	0.405	0.560	1
Major acids	0.646	0.648	0.720	0.753	0.775	0.809	0.856	1

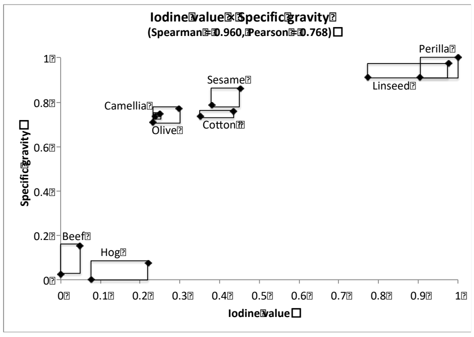


Figure 11 Scatter diagram by two informative features.

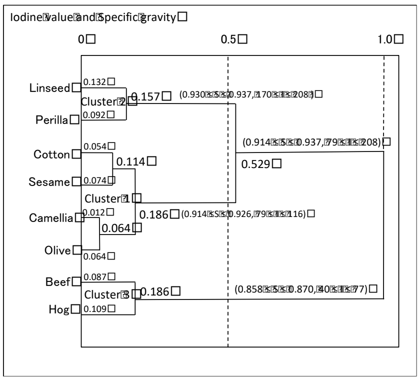

Iodine value and Specific gravity
0
0.5
1.0
Linseed
Perilla
Cotton
Sesame
Camellia
Olive
Beef
Hog
0.132
0.092
0.054
0.074
0.012
0.064
0.087
0.109
Cluster 2
Cluster 1
Cluster 3
0.157
0.114
0.064
0.186
0.186
0.529
(0.930 ≤ S ≤ 0.937, 170 ≤ I ≤ 208)
(0.914 ≤ S ≤ 0.937, 79 ≤ I ≤ 208)
(0.914 ≤ S ≤ 0.926, 79 ≤ I ≤ 116)
(0.858 ≤ S ≤ 0.870, 40 ≤ I ≤ 77)

Figure 12 Descriptions by Specific gravity and Iodine value.

Figure 11 is the scatter diagram of Oil's data for the selected two robustly informative features. This figure shows again three distinct clusters (Linseed, Perilla) and (Cotton, Sesame, Camellia, Olive), and (Beef, Hog). They exist in locally limited regions and they organize again a *geometrically thin structure* with respect to the selected features. Figure 12 shows the dendrogram with concept descriptions of clusters with respect to *Specific gravity* and *Iodine value*. This dendrogram clarifies two major clusters *Plant oils* and *Fats* addition to three distinct clusters, and the compactness take smaller values compared to the dendrogram in Figure 10.

We should note that our exploratory method to analyze distributional data is depending only on the compactness for each feature and combined features. In other words, the measure of feature effectiveness, the measure of similarity between objects and/or clusters, and the measure of cluster quality are based on the same simple notion of the *concept size*.

4.3 Analysis of City temperature data

Guru et al. [10]and several other authors (De Carvalho [11]; De Carvalho and De Souza [12]) used this temperature data in their respective clustering methods. In this data, 12 interval-valued features describe world widely selected 37 cities. The minimum and the maximum

doi:10.20944/preprints202103.0753.v1

temperatures in degree centigrade determine the interval value for each month. We use 0-1 normalized temperatures for each month, and we obtained the PCA result in Figure 13. In this figure, each arrow line connects from the minimum to the maximum quantile vectors and it's length show the concept size. The first principal component has a large contribution ratio and 37 cities line up from *cold* (left) to *hot* (right) in the limited zone between Tehran and Sydney. In this data, we should note that Frankfurt and Zurich have very large concept sizes while Tehran has very small size compared to other cities.

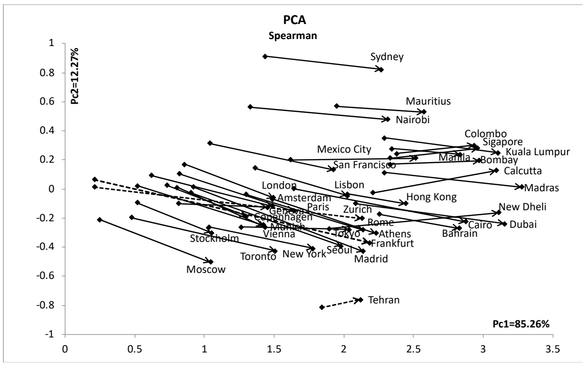


Figure 13 PCA result for City temperature data.

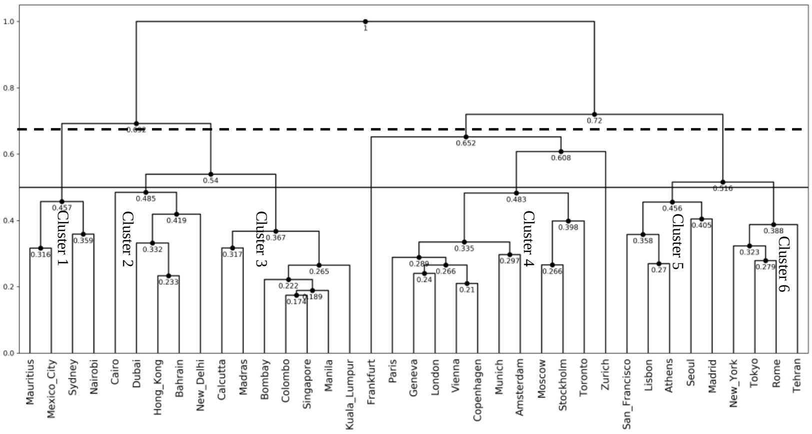


Figure 14 Dendrogram for 12 months.

In Figure 14, we can recognize 6 clusters at cut-point 0.5 excepting Frankfurt and Zurich. De Calvalho and De Souza [12] obtained four clusters by their dynamical clustering methods. We can find exactly the same clusters by cutting our dendrogram as the dotted line in the figure.

Table 7 Average compactness of each feature in selected clustering steps.

Steps	Average compactness for several clustering steps											
	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
0	0.195	0.194	0.224	0.265	0.217	0.281	0.305	0.286	0.289	0.266	0.233	0.194
25	0.360	0.345	0.363	0.406	0.361	0.461	0.519	0.484	0.466	0.422	0.410	0.375
29	0.409	0.389	0.426	0.490	0.414	0.544	0.609	0.555	0.516	0.456	0.443	0.429
31	0.466	0.443	0.476	0.500	0.451	0.593	0.667	0.609	0.568	0.515	0.486	0.476
33	0.489	0.477	0.476	0.500	0.464	0.618	0.694	0.664	0.586	0.522	0.500	0.512
35	0.580	0.568	0.583	0.645	0.656	0.853	0.984	0.969	0.797	0.662	0.608	0.583

Table 7 shows the average values of the compactness for 12 months at selected clustering steps: 25, 29, 31, 33, and 35. The most informative features are *February*, then *January* and *May*. Feature *May* is important to recognize Clusters 1, 2, and 3. The scatter diagram of Figure 15 (a) shows this fact explicitly, where we used the sum of the minimum and the maximum temperatures as feature values. Figure 15(b) is the scatter diagram for *January* and *February*. This figure describes well the mutual relations of Clusters 4, 5, and 6, while the distinctions between Clusters 1, 2, and 3 disappear. We should note that we can reproduce essential structures appearing in the dendrogram by using only *three* selected informative features.

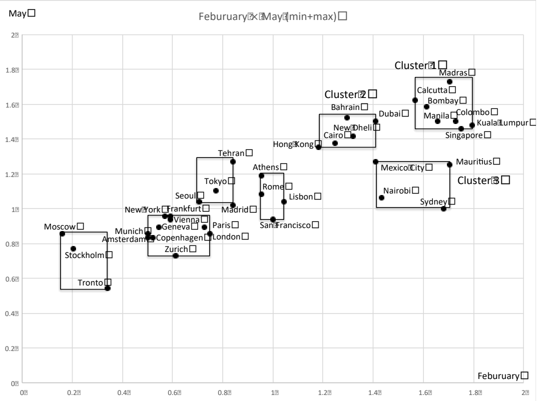


Figure 15 (a) Cluster descriptions for February and May.

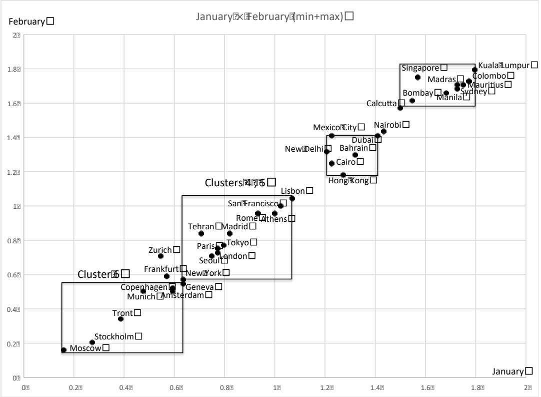


Figure 15 (b) Cluster descriptions for January and February.

4.4 Analysis of the Hardwood data

The data is selected from the US Geological Survey (Climate - Vegetation Atlas of North

America) [15][17][19]. The number of objects is *ten* and the number of features is *eight*. Table 8 shows quantile values for selected *ten hardwoods* under the feature: (Mean) *Annual Temperature* (*ANNT*). We selected the following eight features to describe objects (hardwoods). The data formats for other features F_2~F_8 are the same as in Table 8.

F_1: Annual Temperature (*ANNT*) (ºC); F_2: January Temperature (*JANT*) (ºC);

F_3: July Temperature (*JULT*) (ºC); F_4: Annual Precipitation (*ANNP*) (mm);

F_5: January Precipitation (*JANP*) (mm); F_6: July Precipitation (*JULP*) (mm);

F_7: Growing Degree Days on 5ºC base ×1000 (*GDC*5); and F_8: Moisture Index (*MITM*).

We use the quartile representation by omitting 10% and 90% quantiles from each feature in order to assure the monotone property of our *compactness* measure. As the result, our Hardwood data is a histogram data of the size (10 objects)×(8 features)×(5 quantile values). Table 9 shows a part of our 0-1 normalized Hardwood data, where *five* 8-dimensional quantile vectors describe each hardwood.

Table 8 The original quantile values for ANNT

	Mean Annual Temperature (°C)						
Taxon name	0%	10%	25%	50%	75%	90%	100%
ACER EAST	−2.3	0.6	3.8	9.2	14.4	17.9	23.8
ACER WEST	−3.9	0.2	1.9	4.2	7.5	10.3	20.6
ALNUS EAST	−10.2	−4.4	−2.3	0.6	6.1	15.0	20.9
ALNUS WEST	−12.2	−4.6	−3.0	0.3	3.2	7.6	18.7
FRAXINUS EAST	−2.3	1.4	4.3	8.6	14.1	17.9	23.2
FRAXINUS WEST	2.6	9.4	11.5	17.2	21.2	22.7	24.4
JAGLANS EAST	1.3	6.9	9.1	12.4	15.5	17.6	21.4
JAGLANS WEST	7.3	12.6	14.1	16.3	19.4	22.7	26.6
QUERCUS EAST	−1.5	3.4	6.3	11.2	16.4	19.1	24.2
QUERCUS WEST	−1.5	6.0	9.5	14.6	17.9	19.9	27.2

Table 9 A part of Hardwood data by quantile representation

Taxon name		F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
ACER EAST	0	0.251	0.110	0.165	0.072	0.014	0.124	0.048	0.587
	1	0.406	0.326	0.416	0.163	0.059	0.197	0.167	0.935
	2	0.543	0.452	0.566	0.201	0.102	0.221	0.286	0.967
	3	0.675	0.581	0.700	0.242	0.143	0.250	0.417	0.989
	4	0.914	0.872	0.813	0.336	0.248	0.491	0.798	1.000
ACER WEST	0	0.211	0.124	0.000	0.004	0.006	0.000	0.000	0.065
	1	0.358	0.364	0.213	0.091	0.080	0.051	0.071	0.576
	2	0.416	0.420	0.292	0.145	0.137	0.084	0.119	0.728
	3	0.500	0.518	0.393	0.237	0.263	0.115	0.179	0.902
	4	0.832	0.734	0.828	0.932	0.923	0.354	0.655	1.000

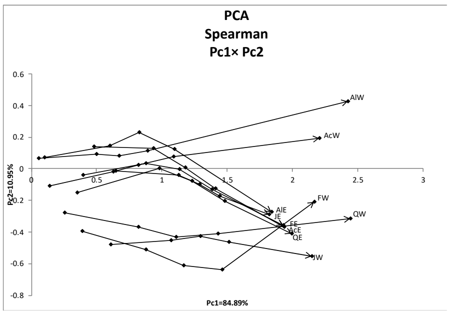


Figure 16 PCA result for Hardwood data.

Figure 16 is the result of PCA by the quantile method. Four line segments connecting from 0% to 100% quantile vectors in the factor plane represent each hardwood. East hard woods have similar shapes, while west hardwoods show significant differences in the last line segments connecting from 75% to 100% quantile vectors. We can recognize three clusters, (Acer West, Alnus West), (five east hard woods), and (Fraxinus West, Juglans West, Quercus West), in this factor plane.

Figure 17 is the result of our HCC for eight features represented by four equiprobability bins. Ten hardwoods, especially AcW and AlW, have very large concept sizes exceeding our simple criterion 0.5. As the result, we have two major chaining clusters: ((((((AcE, JE)FE)QE)AlE)AcW)AlW) and ((FW, JW)QW).

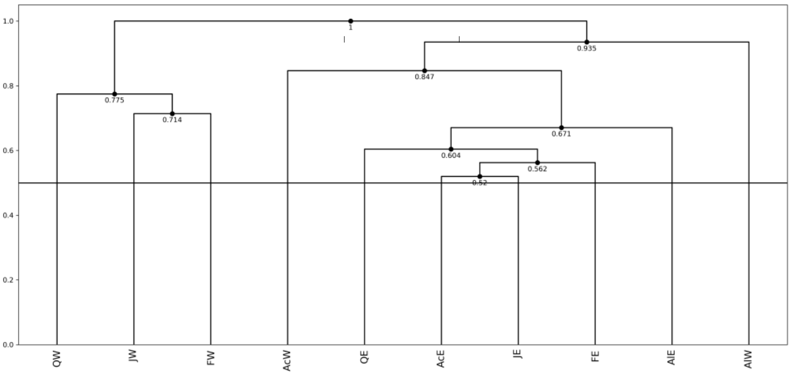


Figure 17 Results of clustering for Hardwoods data (eight features).

Table 10 Average compactness of each feature in each clustering step.

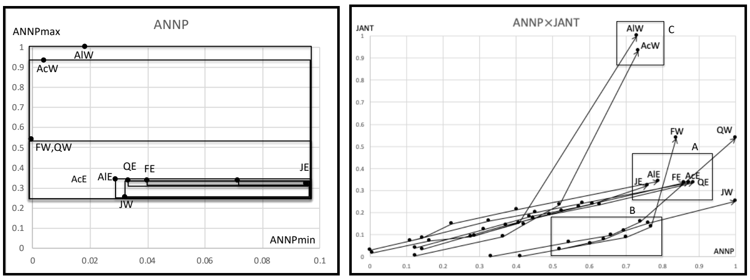
Step	Average compactness of each feature							
	ANNT	JANT	JULT	ANNP	JANP	JULP	GDC5	MITM
0	0.161	0.160	0.178	0.115	0.113	0.133	0.180	0.196
1	0.220	0.228	0.239	0.144	0.140	0.172	0.246	0.242
2	0.229	0.234	0.268	0.186	0.197	0.191	0.256	0.323
3	0.238	0.243	0.282	0.202	0.217	0.203	0.268	0.338
4	0.279	0.269	0.322	0.223	0.243	0.220	0.292	0.358
5	0.404	0.395	0.475	0.337	0.372	0.350	0.455	0.541
6	0.490	0.472	0.570	0.388	0.428	0.401	0.525	0.614
7	0.601	0.578	0.692	0.571	0.595	0.505	0.646	0.739
8	0.829	0.777	0.938	0.768	0.810	0.887	0.899	1.000

Table 10 shows the average compactness of each feature in each clustering step. The most informative feature is *ANNP* then *JULP*. Figure 18(a) shows the nesting structure of rectangles spanned by ten hardwoods with respect to the minimum and the maximum values of *ANNP*. Another representation of the structure is:

((((((((JE,AcE)FE)QE)JW)AlE)(FW,QW))AcW)AlW).

Juglans West is merged to the cluster of east hardwoods. From Figure 18 and Table 10, we see that *JANT* is important to separate the cluster (FW, JW, QW) from the other cluster. In fact, we have the scatter diagram of Figure 18 (b) with respect to *ANNT* and *JANT*. This scatter diagram is very similar to the PCA result in Figure 16. Figure 18 (b) suggests also the cluster descriptions by three rectangles A, B, and C. Rectangles A and C include the maximum quantile vectors of (AcW, AlW) and five east hardwoods, respectively. Rectangle B includes 25%, 50%, and 75% quantile vectors of

(FW, JW, QW). They clarify the distinctions between three clusters.



(a) (b)

Figure 18 Scatter diagrams for the selected informative features.

5. Discussion

This paper presented an exploratory hierarchical method to analyze histogram-valued symbolic data by unsupervised feature seoection. We described each histogram value of each object and/or cluster by predetermined number of equiprobability bins. We defined the notion of the *concept size*, and then the *compactness* for objects and/or clusters as the *measure of similarity* for our hierarchical clustering method. The compactness plays the multi-role measure to evaluate the similarity between objects and/or clusters, to evaluate the dissimilarity of a cluster against the whole concept in each clustering step, and to evaluate the effectiveness of features in the selected clustering steps. We showed the usefulness of the proposed method by four distributional data sets. In each example, we could have *two* or *three* robustly informative features. The scatter diagram and the dendrogram for the selected features reproduced well the essential structures embedded in the given distributional data.

In supervised feature selection for histogram-valued symbolic data, we can use class-conditional hierarchical conceptual clustering addition to our unsupervised feature selection method.

Acknowledgement

This work was supported by JSPS KAKENHI (Grants–in–Aid for Scientific Research) Grant Number 25330268. Part of this work has been conducted under JSPS International Research Fellow program.

References

1. Dy J. G.; Brodley C. E. Feature selection for unsupervised learning, *Journal of Machine Learning Research* 2004, 5, pp.845-889.
2. Liu H.; Motoda H. *Computational Methods of Feature Selection*: CRC Press, London, 2007.
3. Miao J.; Niu L. A survey on feature selection, *Procedia Computer Science* 91, 2016, pp.919-926.
4. Solorio-Fernández S.;Martínez-Trinidad J. F.; Carrasco-Ochoa J. A. A review of unsupervised feature selection methods, *Artificial Intelligence Review* 53, 2020, pp.907-948.
5. Bock H.-H., Diday E. *Analysis of Symbolic Data*; Springer, Berlin, Heidelberg, 2000.
6. Billard L. ; Diday E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Chichester, Wiley, 2007.
7. Diday E. Thinking by classes in data science: the symbolic data analysis paradigm. *WIREs Comput Stat* 2016, 8, pp.172-205. doi: 10.1002/wics.1384.
8. Billard L.; Diday E. *Clustering Methodology for Symbolic Data*, Chichester, Wiley, 2020.

9. Michalski R. S.; Stepp R.(1983) Learning from observation: Conceptual clustering. In: R. S. Michalski, J. G.Carbonell, and T.M. Mitchel, editors, *Machine Learning*, An Artificial Intelligence Approach, Volume II, pages 331-363, Palo Alto, TIOGA Publishing Co.
10. Guru D.S.; KMiranagi B. B.; Nagabushan P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns, *Pattern Recognition* 25, 2004, pp.1203-1213.
11. De Carvalho, F. de A. T. Fuzzy c-means clustering methods for symbolic interval data, *Pattern Recognition Letters*, doi.1016/j.patrec.2006.08.014.
12. De Carvalho, F. de A. T.; De Souza, M.C.R. Unsupervised pattern recognition models for mixed feature-type data, *Pattern Recognition Letters*, 31, 2010, pp.430-443.
13. Ichino M.; Yaguchi H. Generalized Minkowski metrics for mixed feature-type data analysis, *IEEE Trans. Systems, Man, & Cybernetics*, 24, 1994, pp.698-708.
14. Ono Y.; Ichino M. A new feature selection method based on geometrical thickness, , 1, 1998, pp. 19-38.
15. Ichino M. The quantile method of symbolic principal component analysis, *Statistical Analysis and Data Mining*, 4, 2011, pp.184-198.
16. Ichino M.; Brito P. A hierarchical conceptual clustering on the quantile method for mixed feature-type data. IPS079, 59th World Statistics Congress of the International Statistical Institute, 2013, Hong Kong.
17. Ichino M. The data accumulation method: Dimensionality reduction, PCA, and PCA like visualization. *Symbolic Data Workshop*, SDA2015, Orleans, France.
18. Ichino M; and Umbleja K. Similarity and dissimilarity measures for mixed feature-type symbolic data, 48th Scientific Meeting of the Italian Statistical Society SIS2016, Salerno.
19. Histogram data by the U.S. Geological Survey, Climate-Vegetation Atlas of North America. Available: http://pubs.usgs.gov/pp/p1650-b/.