Advances in Data Analysis and Classification (2021) 15:407–436
https://doi.org/10.1007/s11634-020-00411-w

REGULAR ARTICLE

Hierarchical conceptual clustering based on quantile method for identifying microscopic details in distributional data

Kadri Umbleja[1] · **Manabu Ichino**[1] · **Hiroyuki Yaguchi**[1]

Received: 6 December 2018 / Revised: 29 November 2019 / Accepted: 9 July 2020 /
Published online: 22 July 2020

Abstract

Symbolic data is aggregated from bigger traditional datasets in order to hide entry specific details and to enable analysing large amounts of data, like big data, which would otherwise not be possible. Symbolic data may appear in many different but complex forms like intervals and histograms. Identifying patterns and finding similarities between objects is one of the most fundamental tasks of data mining. In order to accurately cluster these sophisticated data types, usual methods are not enough. Throughout the years different approaches have been proposed but they mainly concentrate on the "macroscopic" similarities between objects. Distributional data, for example symbolic data, has been aggregated from sets of large data and thus even the smallest microscopic differences and similarities become extremely important. In this paper a method is proposed for clustering distributional data based on these microscopic similarities by using quantile values. Having multiple points for comparison enables to identify similarities in small sections of distribution while producing more adequate hierarchical concepts. Proposed algorithm, called microscopic hierarchical conceptual clustering, has a monotone property and has been found to produce more adequate conceptual clusters during experimentation. Furthermore, thanks to the usage of quantiles, this algorithm allows us to compare different types of symbolic data easily without any additional complexity.

Keywords Conceptual clustering · Quantile method · Symbolic data

Mathematics Subject Classification 68T09

✉ Kadri Umbleja
kadriumbleja@gmail.com

Manabu Ichino
ichino@mail.dendai.ac.jp

Hiroyuki Yaguchi
yaguchi@mail.dendai.ac.jp

[1] Tokyo Denki University, Saitama, Japan

1 Introduction

Symbolic data approach enables to represent huge amounts of data in compact but complex form. Symbolic data can be considered as an aggregated distribution of a much larger amount of classical data. It summarises a given set of data by concentrating on the main common aspects of original data and hides entry specific details. Despite its aggregated property, the distributional information contains huge amounts of fine details which can be beneficial during data analysis but are mostly overlooked in current methods for analysing symbolic data. These small microscopic details are especially beneficial during clustering when the goal is to find groups with similar objects within and have clear segregation between groups.

In this paper, we offer an approach for hierarchical conceptual clustering which takes into account small details in symbolic data by describing symbolic objects using quantiles. The groups are formed during clustering based on similarities in multiple quantile points between objects. Therefore, the comparison between objects is much more detailed than by just comparing the area objects cover or the start and end points. This proposed method allows to find clusters where objects are similar to each other in microscopic level. Furthermore, the approach ensures that monotone nesting structure is guaranteed during clustering.

In Sect. 2, we cover the previous works in the field of conceptual clustering for symbolic data.

In Sect. 3, we present quantile representation of symbolic data. The quantile method transforms each of n complex symbolic objects to d m-dimensional numeric vectors, called quantile vectors. Quantiles are obtained from underlying distributional information from symbolic object. m presents preselected integer number—larger the m, the more microscopic properties will be considered. If m equals to two, only the start and end point of the symbolic data are used, reducing quantile representation on interval. Therefore, the selection of integral value m controls the granularity of the sub-concepts by constructing the given object in the representation space.

In Sect. 4, we propose an algorithm for hierarchical conceptual clustering based on quantile representation of symbolic data for finding microscopic similarities between compared objects. Then we will use some well known datasets in symbolic data analysis to show the benefits of proposed method and compare results between macroscopic and microscopic approaches.

2 Background

2.1 Symbolic data

The symbolic data analysis (SDA) (Billard and Diday 2006; Diday and Esposito 2003) is an approach to data analysis which permits describing and analysing complex data. If classical data is described by giving a single value to each variable, then the symbolic data appears in many different and complex forms. For example: symbolic data can be an interval, histogram, categorical value or modal valued data. All these different

types can be considered as distributions. This kind of data expands classical data by considering more complete and complex information.

Symbolic data can be extracted from many different sources and in many ways. The most common feature is that bigger traditional data sets are aggregated into more compact forms of data which hides entry specific information and provides a more general and summative overview of the source data. Thanks to these properties, SDA methods enable to analyse large data sets (big data) which are too large to be analysed by usual methods. Furthermore, the fact that aggregated data hides entry specific information makes the symbolic data also suitable for fields where privacy concerns are vital.

2.2 Hierarchical conceptual clustering

The aim of clustering is to form groups (clusters) with objects within a single cluster being similar and those between clusters being dissimilar according to some suitably defined dissimilarity or similarity criteria (Billard and Diday 2006). Many hierarchical clustering methods have been extensively developed.

Conceptual clustering is a paradigm that differs from ordinary data clustering by generating a concept description for each cluster of objects (Michalski and Stepp 1983). Therefore, not only objects with common properties are grouped, but characteristics of each cluster over set of data objects are extracted, describing the regularities in achieved clusters (Hu 1992). This classification scheme can consist of a set of disjointed clusters, or a set of clusters organized into a hierarchy. Each cluster is associated with a generalized conceptual description of the objects within the cluster. Hierarchical clusterings are often described as classification trees (Jonyer et al. 2001).

Conceptual clustering can be used for a variety of tasks, starting from data exploration to model fitting. The aim of applying clustering to a dataset is to gain a better understanding of the data, in many cases by highlighting hierarchical topologies (Jonyer et al. 2001). Clustering aims to achieve clusters that are optimally "connected" or "compact" (Johnson 1967). In Hubert (1972) extensions are offered for Johnson's algorithms. Fisher proposed an incremental hierarchical conceptual clustering method designed to maximize inference ability (Fisher 1987). A comprehensive review of data clustering methods is offered in Jain et al. (1999). Goswami and Chakrabarti (2012) proposed conceptual clustering algorithm based on comparing values of classical dataset against medians and/or quartile values. Their algorithm is applied to classical data with single point values not distributions. Single point is compared with features median/quartiles and labeled according to its position inside the feature. They use combination of labels directly as cluster description with $4d$ possible combinations if using median/quartiles and d is number of features. Therefore, algorithm produces a large number of groups and may require additional merging to avoid cluttering during decision making.

Because symbolic data has a complex structure, it also requires a more complicated approach to conceptual clustering than classical data. A thorough survey of symbolic data clustering methods has been compiled in de Carvalho and de Souza (2010). Generalized Minkowski metrics for mixed feature types based on the Cartesian system

model has been defined in Ichino and Yaguchi (1994). Dendrograms obtained from the application of standard linkage methods are also presented by Ichino and Yaguchi. Different dissimilarity measures for symbolic data clustering are covered in Billard and Diday (2006). SODAS software project produced much research into different aspects of symbolic data clustering like Bertrand and Mufti (2008), Brito and De Carvalho (2008) and De Carvalho et al. (2008). Irpino and Verde (2006) proposed clustering approach based on Wasserstein-Kantorovich metric also using quantile functions. Their difference with current proposed approach to quantiles comes from the aspect that they still use underlying histogram bin values and probabilities. Furthermore, like in Umbleja (2017), the merger of histograms is complex process. Both cases suffer from rising number of quantiles/bins when histograms are merged. The merger has $O(n^3)$ complexity respect to histograms being merged, features and quantiles/bins.

All of those symbolic data conceptual clustering methods are considered to compare "macroscopic" properties of symbolic objects. By "macroscopic" properties, we mean that objects are compared according to general information embedded into their symbolic representation. For example, in Ichino and Umbleja (2018), concept size is used to determine the formation of clusters. Concept size is the area two joined objects span in the Cartesian space. Therefore, basically the start and end value of the feature for an object have impact during clustering. In reality, due to its complex structure and the way larger datasets are aggregated into symbolic form, the symbolic data object contains huge amounts of fine details that are not compared nor considered during this kind of clustering process. Those small fine details are called "microscopic" properties in this paper, reflecting inner structure of symbolic objects.

There has been previous research into conceptual clustering of symbolic data described as quantiles like Brito and Ichino (2010) and Brito and Ichino (2011). The differences between these methods and proposed method is how dissimilarity is measured and how concepts are merged and handled. The similarities are that all methods are comparing distributions on multiple points(quantiles). The main novelty proposed in this paper compared with previous works is the usage of "quantile rectangles" that naturally produce conceptual descriptions for objects and merged concepts.

3 Quantile representation of symbolic data

3.1 Types of symbolic data

Symbolic data can be represented in many different forms. The most common types are interval valued data, histogram valued data, categorical data and modal multi-variable data.

Let set of n objects U be represented by $U = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Let $F_1, F_2, \ldots, F_d$ be d features describing each object ω_i, $i = 1, 2, \ldots, n$ of different with symbolic object E_{ij}.

Definition 1 Let feature F_j be an interval valued feature and let each object $\omega_i \in U$ be represented by an interval:

$$E_{ij} = \{[a_{ij}, b_{ij}]1\} \tag{1}$$

Definition 2 Let feature F_j be a histogram valued feature and let each object $\omega_i \in U$ be represented by a histogram:

$$E_{ij} = \{[a_{ijk}, b_{ijk}]p_{ijk}; k = 1, 2, \ldots, n_{ij}\} \tag{2}$$

where $\sum_{k=1}^{n_{ij}} p_{ijk} = 1$, $b_{ijk} = a_{ij(k+1)}$ and n_{ij} is number of bins for histogram E_{ij}. As can be seen, (1) is a special case of (2) where n_{ij}=1.

Definition 3 Let Y_j be the domain of possible outcomes for modal multi-valued feature F_j containing category values $Y_j = \{c_{j1}, c_{j2}, ..., c_{j|Y_j|}\}$. $|Y_j|$ notates the size of Y_j—the number of category values in Y_j. Each object $\omega_i \in U$ takes categorical values from Y_j with probability p_{ijk} and is represented as:

$$E_{ij} = \{c_{jk}, p_{ijk}; k = 1, 2, \ldots, |Y_j|\} \tag{3}$$

where $\{c_j1, c_j2, ..., c_{j|Y_j|}\} \subseteq Y_j$ and category c_{jk} appears with probability p_{ijk}.

Modal multi-valued data in (3) can be represented as histogram where every category c_{jk} corresponds to bin $[a_{jk}, a_{jk} + l]$ with equal width l and $a_{j(k+1)} = a_{jk} + l$. Therefore, (3) is transformed to histogram as:

$$E_{ij} = \{[a_{jk}, a_{jk} + l]p_{ijk}; k = 1, 2, \ldots, |Y_j|, l = 1/|Y_j|\} \tag{4}$$

The order of categorical values has an impact on the distribution, therefore it should be carefully considered. One possibility is to use sum of frequency and rank values for the ordering (Ichino 2011).

Definition 4 Let Y_j be the domain of possible outcomes for categorical feature F_j containing category values $Y_j = \{c_{j1}, c_{j2}, ..., c_{j|Y_j|}\}$. Each object $\omega_i \in U$ takes categorical values for feature F_j from Y_j and is represented with list of included categories $Y_{ij} \subseteq Y_j$:

$$E_{ij} = \{c_{jk}; c_{jk} \in Y_{ij}\} \tag{5}$$

Categorical data represented as (5) can be represented as modal multi-valued data (3) in following way:

$$E_{ij} = \{c_{jk}, p_{ijk}; k = 1, 2, \ldots, |Y_j|\} \tag{6}$$

where $p_{ijk} = 1$ if category is included in Y_{ij} and $p_{ijk} = 0$ if category is not included in Y_{ij}. Therefore, categorical data (5) can also be represented as histogram data according to (4) as:

$$E_{ij} = \{[a_{jk}, a_{jk} + l]p'_{ijk}; k = 1, 2, \ldots, |Y_j|\} \tag{7}$$

where $p'_{ijk} = p_{ijk}/|Y_{ij}|$ is normalized probability according to the number of categories taken by object ω_i for feature F_j.

It can be seen that different types of symbolic data can be all represented as histograms and therefore considered to be distributions describable with distribution function.

Definition 5 Let an object $\omega_i \in U$ for feature F_j be histogram valued symbolic object as in (2). Therefore object $\omega_i \in U$ for feature F_j can be described with distribution function:

$$F(x) = \begin{cases} 0 & \text{if } x \leq a_{ij1} \\ p_{ij1} \times (x - a_{ij1})/(b_{ij1} - a_{ij1}) & \text{if } a_{ij1} \leq x \leq b_{ij1} = a_{ij2} \\ F(a_{ij2}) + p_{ij2} \times (x - a_{ij2})/(b_{ij2} - a_{ij2}) & \text{if } a_{ij2} \leq x \leq b_{ij2} \\ \dots & \\ F(a_{ijl}) + p_{ijn} \times (x - a_{ijl})/(b_{ijl} - a_{ijl}) & \text{if } a_{ijl} \leq x \leq b_{ijl} \\ 1 & \text{if } b_{ijl} \leq x \end{cases} \tag{8}$$

3.2 Quantile representation

Based on the knowledge of distribution functions, a quantile method (Ichino 2008) offers a common way to represent symbolic data with features of different type. The basic idea is to express the observed feature values by some predefined quantiles of the underlying distribution (Ichino 2011). In case of interval valued data, we assume the uniform distribution inside the interval. In case of histogram valued data, we also assume uniform distribution inside histogram bin.

By using distribution function (8), we can easily obtain m numeric values (quantile values) $Q_1, Q_2, \dots, Q_m$ matching probabilities $p_1, p_2, \dots, p_m$ (where $p_1 < p_2 < \dots < p_m$) in distribution:

$$\begin{aligned} F(Q_1) &= p_1 \\ F(Q_2) &= p_2 \\ &\dots \\ F(Q_m) &= p_m \end{aligned}$$

It should be noted that for quantile values we use term "quantile values" together with minimum and maximum value. If $p_1 = 0$ (minimum value) then $Q_1 = a_{ij1}$ and if $p_m = 1$ (maximum value) then $Q_m = b_{ijl}$. The chosen probabilities may include 0 and 1 but do not have to—in some cases it is beneficial to truncate distributional data.

Definition 6 Let $U = \{\omega_1, \omega_2, \dots, \omega_n\}$ be set of n-objects . Let $F_1, F_2, \dots, F_d$ be d features describing each object ω_i of different feature types. Let m be preselected integer number to determine the common number of quantiles for each feature. Then, object's ω_i feature F_j can be represented with m-tuple quantile vector as:

$$E_{ij} = (Q_{ij1}, Q_{ij2}, \dots, Q_{ijm}) \tag{9}$$

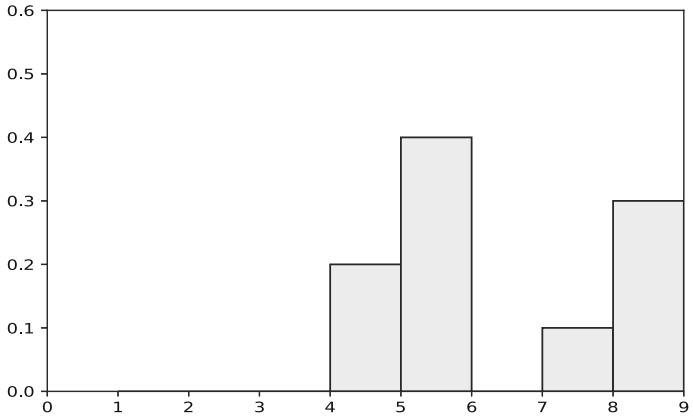

Fig. 1 Histogram for Example 1

where quantiles values $Q_{ij1}, Q_{ij2}, \ldots, Q_{ijm}$ are matching m quantiles $p_1, p_2, \ldots, p_m$ (where $0 \leq p_1 < p_2 < \cdots < p_m \leq 1$).

By default, we can consider quantile value Q_{ijk} to be single points as $Q_{ijk} = \{q_{ijk}\}$. In some cases, as will be shown later, it is beneficial to think of quantile values Q_{ijk} as intervals $Q_{ijk} = [q_{ijkmin}, q_{ijkmax}]$ where $q_{ijkmin} = q_{ijk}$ and $q_{ijkmax} = q_{ijk}$. Therefore, object ω_i becomes m-series of d-dimensional quantile rectangle and (9) can also be written as:

$$E_{ij} = ([Q_{ij1}, Q_{ij1}], [Q_{ij2}, Q_{ij2}], \ldots, [Q_{ijm}, Q_{ijm}]) \tag{10}$$

The advantage of quantile representation of distributional data is that even if underlying symbolic data may be represented by different number of bins (for example in case of histogram—the number of bins n_{ij} may vary from object to object), we can obtain the same number of quantiles for all distributions.

Using m-tuple quantile vectors to describe symbolic objects gives us a simple way how to describe underlying small microscopic details in distribution that would otherwise be overlooked or remain hidden in complex representation of distributional information.

Example 1 Let E_{ij} be histogram described as:

$$E_{ij} = \{[1, 4]0; [4, 5]0.2; [5, 6]0.4; [6, 7]0; [7, 8]0.1; [8, 9]0.3\} \tag{11}$$

The histogram can be seen in Fig. 1. The corresponding distribution function is:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 4 \\ 0.2 \times (x-4) & \text{if } 4 \leq x \leq 5 \\ 0.2 + 0.4 \times (x-5) & \text{if } 5 \leq x \leq 6 \\ 0.6 & \text{if } 6 \leq x \leq 7 \\ 0.6 + 0.1 \times (x-7) & \text{if } 7 \leq x \leq 8 \\ 0.7 + 0.3 \times (x-8) & \text{if } 8 \leq x \leq 9 \\ 1 & \text{if } 9 \leq x \end{cases} \tag{12}$$

We are looking for quantile values Q_1, Q_2,Q_3,Q_4 and Q_5 matching probabilities 0, 0.25, 0.5, 0.75 and 1. We can find these values by simply solving equations $F(Q_1) = 0$, $F(Q_2) = 0.25$, $F(Q_3) = 0.5$, $F(Q_4) = 0.75$ and $F(Q_5) = 1$. We obtain quantiles values Q_1=4, Q_2=5.125, Q_3=5.75, Q_4=8.17 and Q_5=9. Finally, we can produce 5-tuple quantile vector to describe E_{ij}:

$$E_{ij} = (4, 5.125, 5.75, 8.17, 9) \tag{13}$$

We can extend (13) to 5-tuple interval-format quantile vector as (10):

$$E_{ij} = ([4, 4], [5.125, 5.125], [5.75, 5.75], [8.17, 8.17], [9, 9]) \tag{14}$$

4 Microscopic hierarchical conceptual clustering

In the following section, we present the algorithm for Hierarchical Conceptual Clustering (HCC) based of quantile vectors.

We use term "microscopic" for clustering method that takes into account underlying properties from the distribution—for example, to which part of the feature space most of the probabilities are concentrated to? Are the probabilities in compact region or are they spread widely? In the proposed algorithm, the dissimilarity between two objects is measured at multiple points (quantiles) to consider different aspects of distributions. In "macroscopic" approach, only limited characteristics of data are considered—start and end points, join and meet of two objects, for example. In case of intervals where uniform distribution is assumed to be inside a bin, this kind of approach may be adequate, but with more complicated distributions like histograms, many details are overlooked.

4.1 Dissimilarity between two quantile vectors

To compare two quantile vectors $E_{\omega_{xj}}$ and $E_{\omega_{yj}}$ described as m-tuple quantile vectors, the following metric to measure their dissimilarity is proposed.

Definition 7 The dissimilarity between two quantile values Q_{xjk} and Q_{yjk} is:

$$|Q_{xjk} - Q_{yjk}| = max(q_{xjkmax}, q_{yjkmax}) - min(q_{xjkmin}, q_{yjkmin}) \tag{15}$$

where Q_{xjk} is given with interval $[q_{xjkmin}, q_{xjkmax}]$ and Q_{yjk} is given with interval $[q_{yjkmin}, q_{yjkmax}]$.

Definition 8 The dissimilarity between m-tuples quantile vectors $E_{\omega_{xj}}$ and $E_{\omega_{yj}}$ for objects $\omega_x, \omega_y \in U$ for feature F_j is:

$$d_q(E_{\omega_{xj}}, E_{\omega_{yj}}) = \sum_{k=1}^{m} \frac{|Q_{\omega_{xjk}} - Q_{\omega_{yjk}}|}{|D_{jk}|}/m \tag{16}$$

where $|D_{jk}|$ is the length of domain $D_{jk} = [Q_{min_{jk}}, Q_{max_{jk}}]$ for k-th ($k = 1 \ldots m$) quantile over all objects $\omega_i \in U$ for feature F_j.

Normalization among all the quantiles assures that every quantile has an equal impact for the distance between two objects.

Definition 9 The dissimilarity between two objects ω_x and ω_y from set U described with (m)-tuple quantile vectors in d-dimensional feature space is:

$$d(\omega_x, \omega_y) = \sum_{j=1}^{d} \frac{d_q(E_{\omega_{xj}}, E_{\omega_{yj}})}{d} \tag{17}$$

Proposition 1 *The dissimilarity between two quantile vectors $E_{\omega_{xj}}$ and $E_{\omega_{yj}}$ is $0 \leq d_q(E_{\omega_{xj}}, E_{\omega_{yj}}) \leq 1$ as can be seen from (15) to (16).*

Proposition 2 *The dissimilarity between two objects ω_x and ω_y is $0 \leq d(\omega_x, \omega_y) \leq 1$. It is due Proposition 1 and normalization with number of features in (17).*

The measure is basically a join operation between two quantile values in m data points along the distribution function. Therefore, the proposed distance reflects not only the general differences between two comparable quantile vectors but also reflects their inner dissimilarities making it superior over current available methods. The choice of value m and the choice of m probabilities has an impact on result.

Definition 10 Cartesian join of two m-tuple quantile vectors $E_{\omega_{xj}} = (Q_{xj1}, Q_{xj2}, \ldots, Q_{xjm})$ and $E_{\omega_{yj}} = (Q_{yj1}, Q_{yj2}, \ldots, Q_{yjm})$ with respect to the j-th feature is defined as:

$$\begin{aligned} E_{\omega_{xj}} \boxplus E_{\omega_{yj}} = ([&min(Q_{xj1min}, Q_{yj1min}), max(Q_{xj1max}, Q_{yj1max})], \\ &[min(Q_{xj2min}, Q_{yj2min}), max(Q_{xj2max}, Q_{yj2max})], \\ &\ldots, \\ &[min(Q_{xjmmin}, Q_{yjmmin}), max(Q_{xjmmax}, Q_{yjmmax})]) \end{aligned} \tag{18}$$

It can be seen that the result of the Cartesian join of the given two quantile vectors is a vector of intervals. Cartesian join of two objects, described by m-tuple quantile vectors in d-dimensional feature space, is described m-series of d-dimensional rectangles.

The Cartesian join operation is used during HCC when concepts are merged. Lower and upper bounds are taken from both merged concepts at every quantile value point. We can consider the area covered by d-dimensional rectangles as concept size of that specific quantile point.

Example 2 Let E_{pj} be histogram described as quantile vector in (13). Let E_{qj} be described as quantile vector:

$$E_{qj} = (4, 5, 7, 8, 9) \tag{19}$$

We can see that both distributions have the same start and end points. The difference between objects comes from the internal variation in the distribution.

The dissimilarity between E_{pj} and E_{qj} for quantiles $k = 1, \ldots, 5$ is, according to (15) :

$$\begin{aligned}
|Q_{pj1} - Q_{qj1}| &= max(4, 4) - min(4, 4) = 0 \\
|Q_{pj2} - Q_{qj2}| &= max(5, 5.125) - min(5, 5.125) = 0.125 \\
|Q_{pj3} - Q_{qj3}| &= max(7, 5.75) - min(7, 5.75) = 1.25 \\
|Q_{pj4} - Q_{qj4}| &= max(8, 8.17) - min(8, 8.17) = 0.17 \\
|Q_{pj5} - Q_{qj5}| &= max(9, 9) - min(9, 9) = 0
\end{aligned}$$

Assume the domains for 5 quantiles for feature j are: $D_{j1} = [2, 6]$, $D_{j2} = [3, 6]$, $D_{j3} = [4, 8]$, $D_{j4} = [8, 9]$ and $D_{j5} = [8, 12]$. Then, the dissimilarity between quantile vectors E_{pj} and E_{qj} according to (16) is:

$$\begin{aligned}
d_q(E_{pj}, E_{qj}) &= \sum_{k=1}^{5} \frac{|Q_{pjk} - Q_{pjk}|}{|D_{jk}|} / 5 \\
&= \left(\frac{0}{4} + \frac{0.125}{3} + \frac{1.25}{4} + \frac{0.17}{1} + \frac{0}{4} \right) / 5 \\
&= \frac{0.524}{5} = 0.105
\end{aligned} \tag{20}$$

The Cartesian join of quantile vectors E_{pj} and E_{qj} according to (18) is:

$$E_{(p \boxplus q)j} = ([4, 4], [5, 5.125], [5.75, 7], [8, 8.17], [9, 9]) \tag{21}$$

4.2 Algorithm for hierarchical conceptual clustering

1. For each pair of objects ω_i and ω_i' calculate dissimilarity $d(\omega_i, \omega_i')$ as (17). Find a pair of objects ω_p and ω_q that minimize the dissimilarity d.
2. Generate a merged concept ω_{pq} of ω_p and ω_q in U. Delete ω_p and ω_q from U. The new object ω_{pq} (a concept) is described by Cartesian join $E_{pq} = E_p \boxplus E_q$.

3. Repeat step 2 until U contains only one object (the whole concept).

Merged concept in step 2 is Cartesian join of two objects ω_p and ω_q as described in (18).

A notable property of step 2 and step 3 is intentional dual monotone property such that:

The monotone property of the extension:

$$\{\omega_p\} \subseteq \{\omega_p, \omega_q\} \subseteq \{\omega_p, \omega_q, \omega_r\} \subseteq \ldots \tag{22}$$

And the monotone property of dissimilarity:

$$d(\omega_p, \omega_p) \leq d(\omega_p, \omega_q) \leq d(\{\omega_p, \omega_q\}, \omega_r) \leq \ldots \tag{23}$$

From (15), it is clear that dissimilarity (17) is 0 only if the objects are exactly matching. From Proposition 2 it is known that dissimilarity cannot be less than 0. Therefore, first part of (23) holds. For the second part, we use the term "concept size". We define concept size $P(Q_{xjk})$ as:

$$P(Q_{xjk}) = |Q_{xjk}|/|D_{jk}| \tag{24}$$

Concept size of m-dimensional quantile vector is average of quantile values' concept sizes as:

$$P(E_{xj}) = \sum_{k=1}^{m} \frac{P(Q_{xjk})}{m} \tag{25}$$

and concept size of object is average concept size over all features:

$$P(\omega_x) = \sum_{j=1}^{d} \frac{P(E_{xj})}{d} \tag{26}$$

.

Concept size correlates with the area which the object occupies in the Cartesian space and with definitions (15)-(17) . From (18), we can see that when objects are merged, their span in Cartesian space is also merged. If merged concept is further merged, as in $\{\{\omega_p, \omega_q\}, \omega_r\}$, there are two options. In first case, the k-th quantile value for f-th feature of ω_r is inside corresponding quantile value for $\{\omega_p, \omega_q\}$. In this case, both concept size (24) (and also corresponding dissimilarity) remains the same. In second case, the value is outside quantile value for $\{\omega_p, \omega_q\}$. In that case, the merged quantile value will have larger concept size (and larger dissimilarity) than $\{\omega_p, \omega_q\}$. This property will be carried over from quantile value to quantile vector to object level. Therefore, also second part holds. Concept size can also be used to prove extension property (22) the same way.

Nested structure, for example, and tree structure, should be used during bottom up method to memorize the structure of nested objects and corresponding descriptions

by the Cartesian join regions. It should be noted that in each step of the agglomeration process, we not only know the distance between objects, but also know which features create the differences between concepts.

5 Application

To show the benefits of the proposed approach to HCC, it will be applied to well known symbolic datasets and results of proposed microscopic HCC are compared with macroscopic approaches. For macroscopic HCC in case of interval valued datasets, the dissimilarity between objects is found using (Ichino and Brito 2013; Ichino and Umbleja 2018). For histogram valued data, its' extension from Umbleja (2017) is used.

5.1 Oils data

Oils dataset is an interval valued symbolic dataset introduced by Ichino and Yaguchi (1994). It has been excessively used to verify clustering methods for symbolic interval data [for example in Billard and Diday (2006), El-Sonbaty and Ismail (1998), Hardy and Lallemand (2002) and Guru and Nagendraswamy (2006)].

The dataset contains 8 oils. The first six oils are vegetable oils and last two are animal based oils. Pairs of oils (1,2), (3,4) and so on are expected to have similar properties. Vegetable oils 3 to 6 are very similar to each other. The data is given in interval format with fifth feature being categorical variable. Data is normalized according to feature length. Acids are treated as modal multi-valued data with 9 bins correlating to 9 possible acids. Probability 0 or 1 indicates if the category is present or not.

The following five features describe the Oils data:

- F_1: Specific gravity
- F_2: Freezing point
- F_3: Iodine value
- F_4: Saponification value
- F_5: Major fatty acids

First step in microscopic HCC is to decide on proper set of m quantiles to be used. For the current example, following 7 quantiles are used: 0%, 10%, 25%, 50%, 75%, 90% and 100%. As mentioned earlier, the choice of quantiles has an impact on the result. In this case, 7 quantiles have been chosen so that there are many points for comparison. Included are the end and start points which are self-explanatory. 10% and 90% quantiles are useful for ignoring outliers as they truncate the data. Sometimes it may be beneficial to ignore start and end points and use truncation instead. Other three chosen quantiles assure that there are enough points for comparison inside the data.

The second step is finding domains D_{jk} for every feature j and for every quantile k. In current example $j = 1 \ldots 5$ and $k = 1 \ldots 7$. The minimum and maximum values for F_1 for all 7 quantiles can be seen in Table 1.

As domain values are known, dissimilarity between all 8 oils can be found. Results can be seen in Table 2. The smallest dissimilarity is between Cotton and Olive—0.106.

Table 1 Domain values for F_1 in Oils dataset for 7 quantiles

Quantiles:	0%	10%	25%	50%	75%	90%	100%
F_1 min	0.000	0.008	0.019	0.038	0.057	0.068	0.076
F_1 max	0.911	0.920	0.934	0.956	0.978	0.991	1.000

Table 2 Dissimilarity between oils using 7 quantiles after first iteration of microscopic HCC

	Linsead	Perilla	Cotton	Sesame	Camellia	Olive	Beef	Hog
Linsead	0.000	0.297	0.451	0.481	0.449	0.520	0.774	0.854
Perilla		0.000	0.195	0.265	0.358	0.281	0.599	0.597
Cotton			0.000	0.140	0.189	0.106	0.436	0.404
Sesame				0.000	0.182	0.139	0.576	0.462
Camellia					0.000	0.147	0.550	0.503
Olive						0.000	0.436	0.373
Beef							0.000	0.213
Hog								0.000

Table 3 Dissimilarity calculation for Cotton and Olive for F_1

Quantiles	0%	10%	25%	50%	75%	90%	100%
Cotton	0.709	0.715	0.725	0.741	0.756	0.766	0.772
Olive	0.734	0.737	0.741	0.747	0.753	0.757	0.759
Max	0.734	0.737	0.741	0.747	0.756	0.766	0.772
Min	0.709	0.715	0.725	0.741	0.753	0.757	0.759
Max–Min	0.025	0.022	0.016	0.006	0.003	0.009	0.013
$\lvert D_{jk}\rvert$	0.911	0.913	0.915	0.918	0.921	0.923	0.924
(Max–Min)/$\lvert D_{jk}\rvert$	0.028	0.024	0.017	0.007	0.003	0.010	0.014

Dissimilarity calculation for Cotton and Olive for feature 1 over 7 quantiles can be seen in Table 3. The resulting merged concept description can be seen in Table 4.

For macroscopic HCC, a merged concept with smallest average concept size (span) is found. At first iteration, the most similar pair is Olive and Camilla, different from the first pair in microscopic HCC. The merged concept can be seen in Table 5.

Figure 2a represents the result of microscopic HCC and (b) shows results of macroscopic approach to clustering. As can be seen, clear pairs of oils are forming in both dendrograms but the pairs are slightly different. In both graphs the vegetable oils concept (Cotton, Sesame, Camellia and Olive) is very similar. In case of macroscopic (b), the pairs are formed as it was expected from dataset description. In case of microscopic (a), the pairs are formed in different order, reflecting the small differences on how two algorithms compare objects. Based on the properties of microscopic HCC, it can be said that the results in (a) reflect similarity between objects on more than 2 points (start and end) and can therefore be considered more precise. It should be noted that after

Table 4 Concept description for merged concept Cotton–Olive

%	Specific gravity	Freezing point	Iodine	Saponification	Major acids
0	[0.709,0.734]	[0.323,0.415]	[0.232,0.351]	[0.821,0.845]	[0.222,0.222]
10	[0.715,0.737]	[0.331,0.425]	[0.239,0.36]	[0.832,0.856]	[0.267,0.278]
25	[0.725,0.741]	[0.342,0.438]	[0.249,0.372]	[0.848,0.872]	[0.333,0.361]
50	[0.741,0.747]	[0.362,0.462]	[0.265,0.393]	[0.875,0.899]	[0.444,0.5]
75	[0.753,0.756]	[0.381,0.485]	[0.281,0.414]	[0.902,0.926]	[0.556,0.639]
90	[0.757,0.766]	[0.392,0.498]	[0.291,0.426]	[0.918,0.942]	[0.622,0.722]
100	[0.759,0.772]	[0.4,0.508]	[0.298,0.435]	[0.929,0.952]	[0.667,0.778]

Table 5 Concept description for merged concept Olive-Camilla from macroscopic HCC

Specific gravity	Freezing point	Iodine	Saponification	Major acids
[0.709,0.772]	[0.092,0.508]	[0.232,0.298]	[0.821,0.929]	O, P, M, S

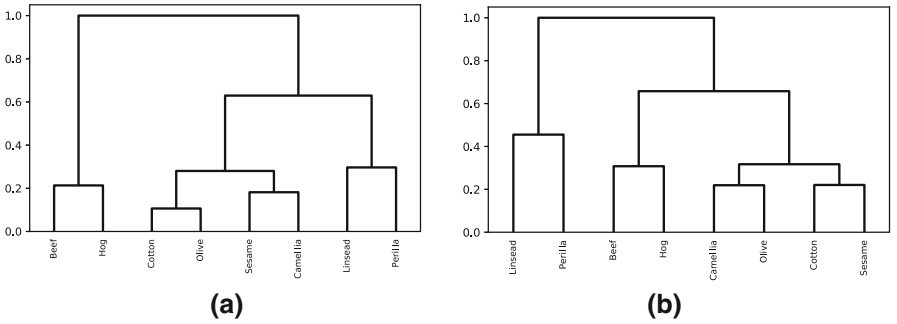

Fig. 2 Dendrograms of microscopic (**a**) and macroscopic (**b**) HCC for Oils dataset

three concepts are left, clusters are matching. The formed seven quantile rectangles at three clusters, that are used in concept description, can be visually followed for features Iodine Value and Specific gravity in Fig. 3.

If dendograms are to be cut at two clusters, clear differences can be followed. In (a) two clusters clearly separate vegetable and animal based oils. In (b), vegetable oils Linsead and Perilla are forming their own cluster while animal fats are clustered together with four similar vegetable oils. It could be argued that result in (a) is more in accordance with real life as the dataset contains two very different kinds of oils. In Fig. 3, it can also be followed that two clusters of vegetable oils look visually closer to each other than to fats. It should also be remembered that Oils data is interval-valued—the proposed microscopic approach does have benefits over macroscopic method but the real advantages become visible with more complex types of symbolic data.

In addition, we can follow cluster quality from dendrogram in Fig. 2. In (a), microscopic case, we can draw the cut dendrogram at 0.28 after 3 clusters are left. In (b), macroscopic case, the dendrogram can be cut after 0.455. Therefore, in (a) the concept

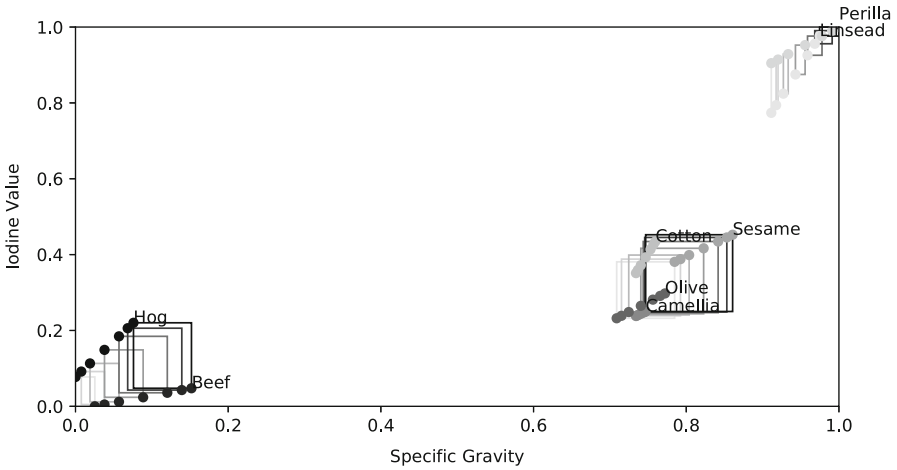

Fig. 3 Quantile values of objects and quantile rectangles at 3 clusters formed during MHCC for Oils dataset features iodine value and specific gravity

size (that correlates to dissimilarity between objects) is smaller and therefore clusters are more compact and more clearly separatable.

5.2 Hardwood data

In this section, a small example using hardwood data by U.S. Geological survey (U.S. Geological Survey 2013) is represented. Ten groups of trees (5 species with west and east coast groups) were chosen with eight features. The features are given with cumulative percentages.
The following eight features describe the Hardwood data:

- F_1: Annual temperature (ANNT) (°C)
- F_2: January temperature (JANT) (°C)
- F_3: July temperature (JULT) (°C)
- F_4: Annual precipitation (ANNP) (mm)
- F_5: January precipitation (JANP) (mm)
- F_6: July precipitation (JULP) (mm)
- F_7: Growing degree days on 5°C base * 1000 (GDC5)
- F_8: Moisture index (MITM)

The hardwood data is in a histogram format—therefore span and concept size of the object are not detailed enough for adequate comparison. All objects tend to cover most of the feature space and therefore have a similar span. Dissimilarities between objects are hidden in their distributions. The quantile values for F_1 can be seen in Table 6.

The structure of Hardwood data can be seen in Fig. 4. The distributional data is represented by using accumulated quantile values to show the fine details of distribution (Ichino and Britto 2014). 7 monotone lines containing 8 points are used to visualize the accumulation of values for quantiles k, $k = 1, \ldots, 7$ over features j, $j = 1, \ldots, 8$. Different grey tones are used for different features. The shapes of those monotone

Table 6 Quantile values for annual temperature

Quantiles	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
Juglans East	1.3	6.9	9.1	12.4	15.5	17.6	21.4
Juglans 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

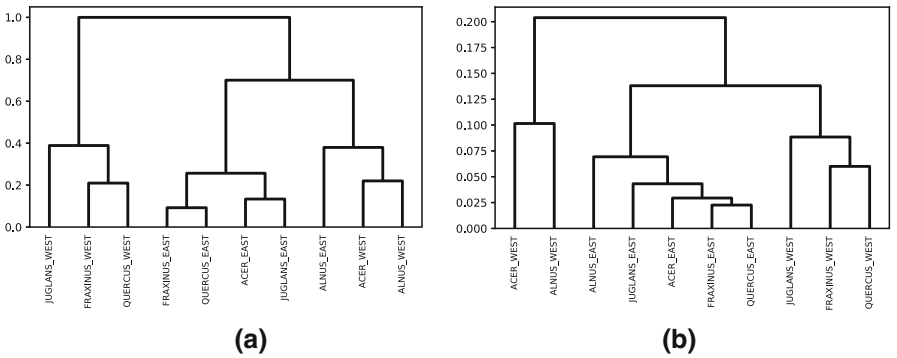

Fig. 4 Accumulation of quantiles over features for Hardwood data

lines allows to identify the patterns in data. For example, for all tree species the difference between east and west lays between the last two quantile values in quantile vector as bottom parts of the figure have similar shapes for the respective species. It can also be seen in Fig. 4 that all east hardwood, except Alnus East, are very similar. All that information should be taken account during HCC. The dendrograms achieved by microscopic and macroscopic HCC can be seen in Fig. 5. Same 7 quantiles as in case of Oils data were used. As noted previously, for histogram-valued data, extension of Ichino and Umbleja (2018) method is used. The original method is not suitable for histogram-valued data as it calculates join and meet by the span the data covers in feature space. Histograms tend to span over large parts of feature space but the regions with high probabilities are important. The extension of concept size method does consider the shape of histogram. Therefore the concept size and dissimilarity will not equal to 1 as would be expected in a dendrogram. It only occurs when histogram has one bin (interval). Despite that drawback, the propositions of dendrogram and concepts merged can be normalized and are are comparable.

There are some fundamental differences between (a) and (b) in Fig. 5 due to underlying distributions as in Fig. 4. In both cases 3 clear clusters emerge but they are

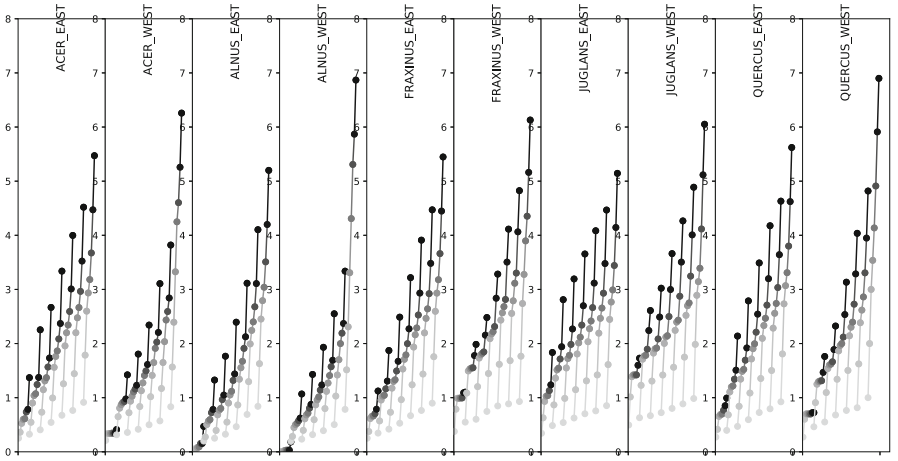

Fig. 5 Dendrograms of microscopic (**a**) and macroscopic (**b**) HCC for Hardwood

not identical. In (b) clusters are formed following real-life expectations—east and west coast trees are grouped independently. In (a) Alnus East gets mixed with Acer and Alnus West. In (b) span has huge impact on compactness measure—dissimilarity between Acer and Alnus West is rather large. In (a), dissimilarity is measured along multiple points and inner distribution of histograms is analysed. Those three trees actually have very similar patterns from Fig. 4. The concept descriptions at 3 clusters for microscopic approach can be seen in Table 7.

It can be further verified with principal component analysis in Fig. 6. The main difference between East and West coast comes from last quantile—in case of west trees, the difference between last and penultimate quantile covers a very large span—meaning that there is a low concentration of probabilities. In (a), due to the normalization factor where differences among quantiles have equal width, the dissimilarity along the last quantile does not overwrite similarities among other 6 quantiles used. Therefore, (a) offers us benefits over (b) by presenting new information that was not known with background knowledge about the dataset but this can be verified with other methods. Furthermore, it makes sure that differences in minimum and maximum values, that may have only low probabilities due to the fact that natural data has normal distribution, due not overwrite other similarities.

When comparing cluster quality, it can be visually followed that if we cut dendrograms in Fig. 5 when 3 clusters remain (the most optimal point), the formed concepts in (a) have better compactness than in (b). Formal comparison of cluster quality between formed microscopic and macroscopic HCC can be followed in Table 8. Three measures where chosen: Calinski–Harabasz (CH), Davies–Bouldin (DB) and Root-mean-square standard deviation (RMSSTD). CH index uses average between- and within-cluster sum of squares. Larger the value, better the cluster. DB also uses within-group and between-group distances to validate formed clusters. Smaller the value, better the cluster. RMSSTD considers only homogeneity within the cluster (Liu et al. 2010; Vendramin et al. 2010).

Table 7 Concept descriptions for 3 clusters formed with MHCC

	0%	10%	25%	50%	75%	90%	100%
	((ACER_EAST-JUGLANS_EAST)-(FRAXINUS_EAST-QUERCUS_EAST))						
F1	[0.25–0.34]	[0.32–0.48]	[0.41–0.54]	[0.53–0.62]	[0.67–0.73]	[0.76–0.79]	[0.85–0.92]
F2	[0.11–0.29]	[0.22–0.38]	[0.31–0.45]	[0.44–0.52]	[0.57–0.63]	[0.66–0.71]	[0.76–0.88]
F3	[0.16–0.3]	[0.36–0.49]	[0.42–0.56]	[0.57–0.66]	[0.7–0.74]	[0.76–0.78]	[0.81–0.93]
F4	[0.03–0.1]	[0.07–0.15]	[0.12–0.17]	[0.18–0.21]	[0.23–0.24]	[0.27–0.28]	[0.32–0.34]
F5	[0.01–0.01]	[0.02–0.03]	[0.03–0.06]	[0.08–0.11]	[0.13–0.14]	[0.18–0.19]	[0.22–0.25]
F6	[0.04–0.12]	[0.12–0.17]	[0.16–0.2]	[0.21–0.22]	[0.24–0.25]	[0.27–0.3]	[0.45–0.49]
F7	[0.05–0.11]	[0.13–0.23]	[0.17–0.29]	[0.27–0.36]	[0.4–0.49]	[0.55–0.61]	[0.7–0.82]
F8	[0.14–0.6]	[0.57–0.88]	[0.82–0.93]	[0.95–0.97]	[0.97–0.99]	[0.99–1.0]	[1.0–1.0]
	((ACER_WEST-ALNUS_WEST)-ALNUS_EAST)						
F1	[0.0–0.21]	[0.19–0.31]	[0.23–0.36]	[0.32–0.42]	[0.39–0.5]	[0.5–0.69]	[0.78–0.84]
F2	[0.0–0.12]	[0.08–0.33]	[0.14–0.36]	[0.22–0.42]	[0.4–0.52]	[0.53–0.61]	[0.73–0.79]
F3	[0.0–0.0]	[0.16–0.23]	[0.21–0.29]	[0.27–0.35]	[0.32–0.48]	[0.39–0.7]	[0.79–0.83]
F4	[0.0–0.03]	[0.05–0.06]	[0.07–0.09]	[0.09–0.15]	[0.15–0.24]	[0.25–0.39]	[0.34–1.0]
F5	[0.0–0.01]	[0.03–0.04]	[0.03–0.08]	[0.05–0.14]	[0.12–0.26]	[0.16–0.4]	[0.25–1.0]
F6	[0.0–0.06]	[0.02–0.13]	[0.05–0.16]	[0.08–0.2]	[0.12–0.24]	[0.16–0.28]	[0.35–1.0]
F7	[0.0–0.0]	[0.05–0.06]	[0.07–0.08]	[0.1–0.12]	[0.12–0.21]	[0.18–0.43]	[0.56–0.69]
F8	[0.07–0.32]	[0.45–0.54]	[0.55–0.72]	[0.7–0.96]	[0.86–0.99]	[0.97–1.0]	[1.0–1.0]
	((FRAXINUS_WEST-QUERCUS_WEST)-JUGLANS_WEST)						
F1	[0.27–0.49]	[0.46–0.63]	[0.55–0.67]	[0.68–0.75]	[0.76–0.85]	[0.81–0.89]	[0.93–1.0]
F2	[0.33–0.52]	[0.45–0.6]	[0.54–0.64]	[0.66–0.7]	[0.74–0.77]	[0.8–0.86]	[0.84–1.0]
F3	[0.1–0.37]	[0.34–0.48]	[0.44–0.53]	[0.52–0.64]	[0.64–0.82]	[0.76–0.87]	[0.91–1.0]
F4	[0.0–0.03]	[0.04–0.06]	[0.06–0.08]	[0.09–0.12]	[0.13–0.16]	[0.18–0.23]	[0.25–0.54]
F5	[0.0–0.01]	[0.01–0.02]	[0.02–0.03]	[0.02–0.04]	[0.03–0.11]	[0.05–0.32]	[0.25–0.62]
F6	[0.0–0.0]	[0.0–0.11]	[0.03–0.17]	[0.1–0.35]	[0.13–0.44]	[0.19–0.5]	[0.46–0.77]
F7	[0.02–0.18]	[0.15–0.35]	[0.23–0.4]	[0.42–0.52]	[0.56–0.69]	[0.64–0.76]	[0.81–1.0]
F8	[0.0–0.13]	[0.21–0.37]	[0.33–0.53]	[0.45–0.66]	[0.61–0.76]	[0.76–0.88]	[0.93–0.99]

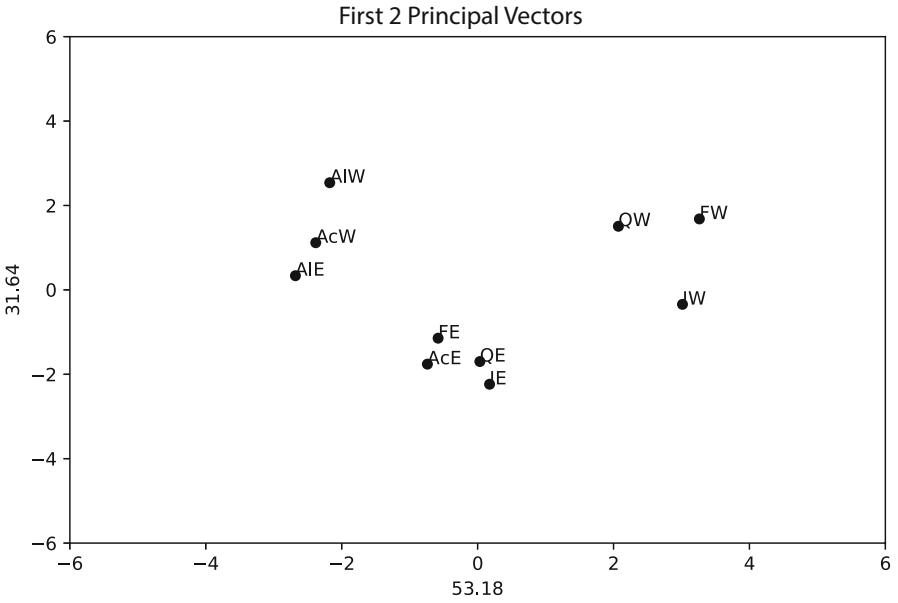

Fig. 6 PCA results of Hardwood data using Centers method (Billard and Diday 2006)

Most of the indexes for cluster validation consider two aspects—compactness and separation (Liu et al. 2010). As symbolic data is not point value but distribution that covers large part of the feature space (that is especially the case with American States' data), there exists lot of overlap between objects and formed clusters do not produce clear separation over whole distribution/span. Based on distance calculation during microscopic HCC, most similar objects over majority of quantiles are merged into cluster. Concept description is achieved at quantile level (as could be followed from Table 7). Separation over one quantile is adequate for achieving separation between clusters. RMSSTD was chosen as quality measure as it only considers compactness of the formed clusters that is important for distributions.

Cluster quality measures are calculated at quantile level and averaged over all quantiles. Even better results can be achieved when instead of average best quantile is used. Calinski–Harabasz index is calculated based on Vendramin et al. (2010) with (27) where $trace(W_q)$ and $trace(B_q)$ are defined as in (28) and (29).

$$\max_{q\in[Q_1\cdots Q_m]} \frac{trace(B_q)}{trace(W_q)} \times \frac{N-k}{k-1} \tag{27}$$

$$trace(W_q) = \sum_{l=1}^{k}\sum_{j=1}^{d}\sum_{\omega_i\in||C_l||}(x_{ijq}-\overline{x}_{ljq})^2 \tag{28}$$

$$trace(B_q) = \sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ijq}-\overline{x}_{jq})^2 - trace(W_q) \tag{29}$$

Table 8 Cluster quality evaluations for Hardwood data

k	Calinski–Harabasz			Davies–Bouldin			RMSSTD		
	MHCC			MHCC			MHCC		
	Max	Avg	HCC	Max	Avg	HCC	Max	Avg	HCC
5	20.656	11.474	5.572	0.201	0.407	0.440	0.043	0.061	0.096
4	16.954	10.522	6.833	0.258	0.451	0.523	0.053	0.073	0.105
3	21.569	13.054	10.872	0.288	0.454	0.708	0.055	0.077	0.115
2	9.027	6.484	3.511	0.350	0.701	0.985	0.093	0.109	0.163

Davies–Bouldin for specific quantile is calculated as generally with distances being calculated as described previously (17). Then, best or average value over all quantiles is used. Same applies with root-mean-square standard deviation.

As can be seen in Table 8, with Calinski–Harabasz measure, 3 clusters are clearly the best result with all different approaches. With Davies–Bouldin, the results are not so clear with index value dropping with larger number of clusters. It can be followed from dendrogram at Fig. 5 that with more than 3 clusters, individual objects are starting to form solo clusters that have good compactness (0 as only one object in cluster) and separation. RMSSTD best value is defined as "elbow" (Liu et al. 2010) and in all cases, it is clearly visible that there is large drop on values before 3 clusters but the index does not change that much with larger number of clusters.

5.3 American States' weather

In this section, the results of proposed methodology are described on climate dataset (National Climatic Data Center 2014). This dataset contains sequential monthly "time bias corrected" average temperature data for 48 states of USA (Alaska and Hawaii are not represented in the dataset). Time period of 1895–2009 is used for comparison purposes. The data provided was first transformed into histograms describing average temperature for every state and every month.
The following twelve features describe the American States data:

- F_1: January
- F_2: February
- F_3: March
- F_4: April
- F_5: May
- F_6: June
- F_7: July
- F_8: August
- F_9: September
- F_{10}: October
- F_{11}: November
- F_{12}: December

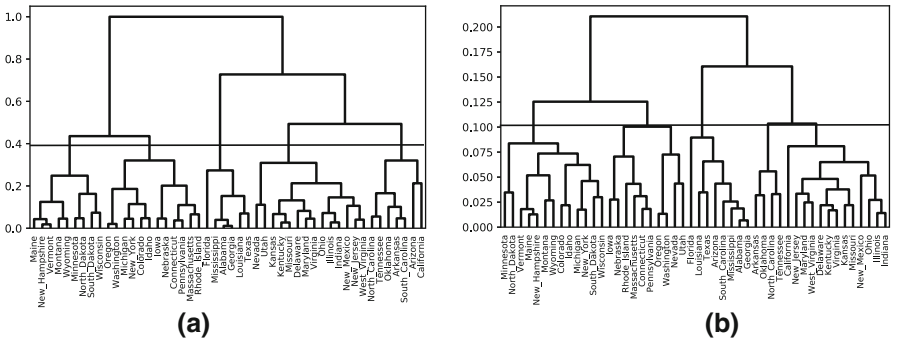

Fig. 7 Dendrograms of microscopic (**a**) and macroscopic (**b**) HCC for states data

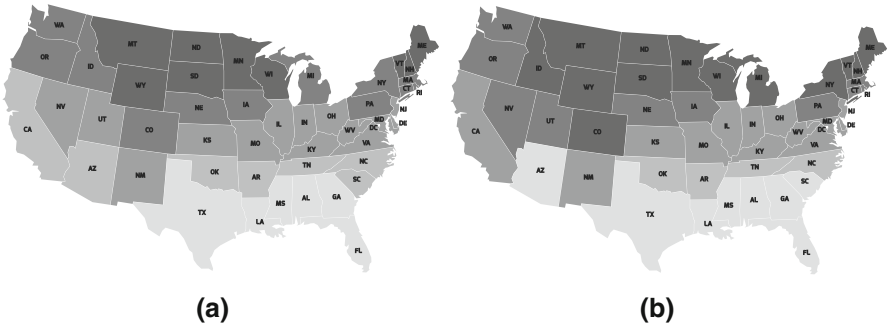

Fig. 8 Five clusters achieved in microscopic (**a**) and macroscopic (**b**) HCC for states data on a map

Table 9 Differences in clustering between Microscopic and Macroscopic HCC—extreme summers

State	MicroHCC	MacroHCC	Avg summer	Extr. summer	Avg in group
Arizona	Warm	Warmest	30	38	30

The dendrograms achieved by microscopic and macroscopic HCC can be seen in Fig. 7. Dendrograms were cut at 5 clusters and plotted to map as can be seen in Fig. 8. Same 7 quantiles as before were used.

As can be seen from Fig. 7, dendograms have different structure. It can be said that (a) would produce 5 compact clusters and (b) would produce 4 clusters. For comparison purposes, 5 clusters are chosen for Fig. 8. We can think of these clusters as "very warm", "warm", "mild","colder" and "cold". Both (a) and (b) in Fig. 8 are reasonable clusters as they follow expected real-life knowledge—clusters follow latitudes with coastal states being clustered to warmer latitude clusters. Inland states are usually clustered with colder groups than others in the same latitude.

It can be followed from Fig. 7 that in both cases the "cold" and "colder" clusters are forming their own branch. Furthermore, it should be noted that Florida has a very different weather pattern from the rest of very warm states by having extremely warm

Table 10 Differences in clustering between Microscopic and Macroscopic HCC—extreme winters

State	MicroHCC	MacroHCC	Avg winter	Extr. winter	Avg in group
California	Warm	Mild	6	− 12	1
Colorado	Cold	Coldest	− 6	− 18	− 8
Michigan	Cold	Coldest	− 9	− 23	− 8
New York	Cold	Coldest	− 9	− 18	− 8
Nevada	Mild	Cold	− 3	− 18	− 5
Utah	Mild	Cold	− 5	− 23	− 5

winters. That property can be followed in dendrograms as Florida is an outlier which is being grouped as an individual object.

If comparing (a) and (b) in Fig. 7, the clusters tend to be more compact in (a) than in (b) meaning better cluster quality. The concept descriptions of five formed clusters using microscopic approach can be seen in Table 11.

Microscopic approach follows the average weather patterns while occurrences of extreme maximums and minimums have a huge impact on macroscopic approach. The differences in clusters can be ascribed to that characteristic. The states clustered to different groups are listed in Tables 9 and 10. States average and extreme temperatures are given. For comparison, cluster's average temperature, according to microscopic HCC, is also included. It can be followed that state's average temperature is similar to cluster's average. The extreme temperature is very different from both state's and cluster's average.

In all cases listed in Tables 9 and 10, the differences are caused by extreme recorded temperatures that have a low probabilty of occurring. In case of macroscopic HCC, those extremes (minimum and maximum temperatures) have a very strong impact on clustering results. In case of microscopic clustering when 7 quantile values were used, the extremes do not have such a huge impact—they are only 2 quantile values and all quantiles' distances are treated equally. Therefore, if 5 out of 7 quantile values are very similar, they have a larger impact on overall similarity than 1 or 2 large differences with minimum and maximum extremes would have. The average weather pattern has more impact than the rare occurrence of few unusual temperatures.

Arizona's concept description can be seen in Table 12. In microscopic HCC, Arizona is grouped with warmer states while in macroscopic approach it groups with the warmest. When comparing Table 12 with warm cluster description in Table 11, it can be followed that for most winter months and most quantiles, Arizona affects the maximum value. For example, for January for quantiles 10–100%, Arizona's quantile value is the maximum for cluster's description. The same Arizona's January quantile values would be minimum values for "Very Warm" cluster showing clearly that Arizona is a borderline case. The low probability high temperatures during summer (for example 100% quantiles for F6–F9) are actually higher than in "Very Warm" cluster. For June (F6), 100% quantile has value 38. 90%, on the other hand, has value 31, meaning 10% of Arizona's distribution for June is covered by 7 degrees. The value at 75% quantile is 28 and therefore 15% of distribution is covered with only 3 degrees showing that high

Table 11 Concept descriptions in Celsius for 5 clusters formed with microscopic HCC

	0%	10%	25%	50%	75%	90%	100%
	Very warm—containing Florida, Texas etc						
F1	[− 7 to 4]	[2—11]	[5–13]	[7–16]	[9–19]	[13–21]	[21–27]
F2	[− 7 to 4]	[5–11]	[6–14]	[8–17]	[11–20]	[14–22]	[21–32]
F3	[− 1 to 10]	[8–15]	[11–17]	[13–19]	[15–22]	[18–25]	[21–27]
F4	[10–16]	[12–17]	[16–19]	[18–22]	[19–24]	[20–26]	[27–32]
F5	[10–21]	[17–22]	[19–23]	[22–24]	[24–26]	[26–27]	[27–32]
F6	[16–21]	[22–22]	[23–24]	[24–27]	[26–30]	[29–31]	[32–32]
F7	[21–21]	[22–27]	[24–28]	[27–29]	[29–31]	[31–32]	[32–38]
F8	[21–21]	[22–27]	[23–28]	[26–29]	[29–31]	[31–32]	[32–38]
F9	[16–21]	[20–22]	[22–24]	[24–28]	[25–30]	[26–31]	[32–32]
F10	[10–16]	[13–19]	[16–22]	[18–23]	[20–25]	[21–26]	[27–32]
F11	[− 1 to 10]	[7–14]	[10–17]	[12–20]	[14–23]	[15–25]	[21–32]
F12	[− 7 to 4]	[5–11]	[6–14]	[8–17]	[10–20]	[13–23]	[21–27]
	Warmer—containing Arkansas, South Carolina etc						
F1	[− 12 to − 7]	[− 1 to 1]	[0–5]	[2–7]	[5–10]	[8–13]	[10–21]
F2	[− 7 to − 1]	[0–3]	[2–6]	[5–9]	[7–12]	[9–14]	[16–21]
F3	[− 7 to − 1]	[3–7]	[7–10]	[9–12]	[12–15]	[14–18]	[21–27]
F4	[− 1 to 10]	[7–12]	[10–14]	[13–17]	[15–19]	[17–21]	[21–27]
F5	[4–10]	[10–17]	[12–18]	[15–21]	[19–24]	[21–26]	[27–32]
F6	[10–16]	[14–22]	[17–23]	[19–25]	[21–28]	[25–31]	[32–38]
F7	[10–21]	[17–23]	[19–26]	[22–28]	[25–31]	[28–35]	[32–38]
F8	[10–21]	[17–23]	[18–25]	[22–28]	[25–30]	[26–33]	[32–38]
F9	[4–16]	[14–19]	[17–22]	[19–24]	[22–27]	[25–30]	[27–38]
F10	[4–10]	[9–12]	[12–16]	[14–18]	[17–21]	[20–24]	[21–32]
F11	[− 7 to − 1]	[3–7]	[6–10]	[8–12]	[10–14]	[13–17]	[16–21]
F12	[− 7 to − 7]	[0–3]	[1–5]	[4–7]	[7–10]	[9–13]	[10–16]

temperatures are not that likely. During clustering process, all quantiles and all months are considered. In majority of cases, especially with lower quantiles and summer months, Arizona is clearly colder than "Very Warm" cluster description and therefore clustering result by microscopic approach is more accurate than result achieved with macroscopic.

Similar phenomena can be followed with California, as can be followed in Table 13, but with winters. California has recorded some very cold winter temperatures but on average, its weather all year round is quite even with small difference between 25 and 75% quantile values for all months (oceanic climate). That makes it different from the average state's weather in "warm" cluster where summers are usually much warmer than in California. When considering other seasons, the average weather in California is very similar to rest of "warm" weather states.

Table 11 continued

	0%	10%	25%	50%	75%	90%	100%
	Mild—containing Kansas, New Mexico etc						
F1	[− 23 to − 7]	[− 10 to − 3]	[− 6 to 0]	[− 4 to 2]	[0–4]	[3–7]	[10–16]
F2	[− 18 to − 7]	[− 6 to − 1]	[− 5 to 0]	[− 2 to 3]	[1–6]	[3–10]	[10–16]
F3	[− 12 to − 1]	[− 1 to 4]	[1–6]	[3–7]	[7–10]	[9–13]	[16–21]
F4	[− 1 to 4]	[5–10]	[6–11]	[8–13]	[11–15]	[14–18]	[16–21]
F5	[− 1 to 10]	[9–16]	[11–17]	[13–18]	[16–20]	[19–22]	[21–32]
F6	[4–16]	[13–20]	[16–22]	[19–24]	[21–25]	[24–27]	[27–32]
F7	[10–21]	[17–22]	[20–23]	[23–25]	[25–28]	[26–30]	[27–38]
F8	[10–16]	[17–22]	[19–23]	[22–24]	[24–27]	[26–30]	[27–38]
F9	[4–16]	[11–16]	[14–18]	[17–20]	[20–23]	[21–25]	[27–32]
F10	[− 1 to 10]	[5–11]	[7–12]	[10–13]	[13–15]	[15–19]	[21–27]
F11	[− 7 to − 1]	[− 1 to 5]	[0–6]	[3–8]	[5–9]	[9–12]	[16–16]
F12	[− 18 to − 7]	[− 8 to − 1]	[− 5 to 0]	[-3 to 3]	[1–5]	[4–8]	[10–16]
	Cold—containing Michigan, Connecticut etc						
F1	[− 23 to − 12]	[− 13 to -6]	[− 11 to − 3]	[− 7 to 1]	[− 4 to 4]	[− 2 to 7]	[4–10]
F2	[− 23 to − 12]	[− 11 to − 3]	[− 9 to 0]	[− 6 to 3]	[− 3 to 6]	[− 2 to 8]	[4–16]
F3	[− 12 to − 7]	[− 6 to 0]	[− 4 to 3]	[− 1 to 6]	[2–8]	[4–9]	[10–16]
F4	[− 7 to 4]	[1–5]	[5–7]	[6–9]	[8–12]	[10–14]	[16–16]
F5	[− 1 to 4]	[6–11]	[9–13]	[12–15]	[14–18]	[15–20]	[21–27]
F6	[4–16]	[11–17]	[12–18]	[15–21]	[18–24]	[20–25]	[21–27]
F7	[10–16]	[13–21]	[16–22]	[18–24]	[20–26]	[23–27]	[27–32]
F8	[10–16]	[14–19]	[16–22]	[18–23]	[20–25]	[21–26]	[27–32]
F9	[4–10]	[11–16]	[12–17]	[14–18]	[16–20]	[19–21]	[21–27]
F10	[− 1 to 4]	[5–10]	[6–11]	[8–12]	[10–14]	[13–15]	[16–21]
F11	[− 12 to − 1]	[− 4 to 4]	[− 1 to 5]	[1–7]	[3–8]	[4–9]	[10–16]
F12	[− 18 to − 7]	[− 10 to − 4]	[− 7 to − 1]	[− 5 to 1]	[− 2 to 4]	[0–7]	[4–10]
	Coldest—containing New Hampshire, Montana etc						
F1	[− 29 to − 18]	[− 21 to − 12]	[− 17 to − 10]	[− 14 to − 7]	[− 10 to − 4]	[− 7 to − 2]	[− 1 to 4]
F2	[− 29 to − 18]	[− 17 to − 11]	[− 15 to − 8]	[− 11 to − 5]	[− 8 to − 2]	[− 4 to 1]	[4–10]
F3	[− 18 to − 12]	[− 10 to − 6]	[− 7 to − 4]	[− 4 to 0]	[− 1 to 2]	[2–4]	[4–16]
F4	[− 7 to − 1]	[0–3]	[2–5]	[5–7]	[7–9]	[9–10]	[10–16]
F5	[− 1 to 4]	[5–10]	[7–11]	[10–13]	[13–15]	[14–17]	[16–27]
F6	[4–10]	[11–16]	[12–17]	[15–18]	[18–20]	[20–22]	[21–27]
F7	[10–16]	[14–18]	[17–21]	[19–23]	[20–25]	[21–26]	[27–32]
F8	[10–16]	[13–17]	[16–19]	[18–21]	[20–24]	[21–26]	[27–27]
F9	[4–10]	[8–11]	[11–13]	[13–16]	[15–19]	[15–20]	[21–27]
F10	[− 7 to − 1]	[2–5]	[5–6]	[7–8]	[8–11]	[10–14]	[16–16]
F11	[− 18 to − 7]	[− 8 to − 1]	[− 5 to 0]	[− 3 to 2]	[0–3]	[3–4]	[10–10]
F12	[− 23 to − 18]	[− 16 to − 11]	[− 14 to − 8]	[− 10 to − 5]	[− 8 to − 3]	[− 4 to − 1]	[− 1 to 4]

Table 12 Concept descriptions for Arizona using 7 quantiles

	0%	10%	25%	50%	75%	90%	100%
F1	[−12]	[0]	[4]	[7]	[10]	[13]	[21]
F2	[−7]	[2]	[6]	[9]	[12]	[14]	[21]
F3	[−1]	[6]	[8]	[12]	[15]	[18]	[27]
F4	[4]	[10]	[12]	[15]	[19]	[21]	[27]
F5	[4]	[14]	[17]	[20]	[23]	[26]	[32]
F6	[16]	[19]	[22]	[25]	[28]	[31]	[38]
F7	[16]	[22]	[25]	[28]	[31]	[35]	[38]
F8	[16]	[21]	[23]	[26]	[30]	[33]	[38]
F9	[10]	[18]	[21]	[24]	[27]	[30]	[38]
F10	[4]	[12]	[16]	[18]	[21]	[24]	[32]
F11	[−1]	[5]	[8]	[12]	[14]	[17]	[21]
F12	[−7]	[1]	[5]	[7]	[10]	[13]	[16]

Table 13 Concept descriptions for California using 7 quantiles

	0%	10%	25%	50%	75%	90%	100%
F1	[−12]	[−1]	[5]	[7]	[9]	[12]	[16]
F2	[−7]	[1]	[6]	[8]	[12]	[14]	[16]
F3	[−7]	[3]	[7]	[11]	[13]	[15]	[21]
F4	[−1]	[7]	[10]	[13]	[15]	[17]	[27]
F5	[4]	[10]	[12]	[15]	[19]	[21]	[27]
F6	[10]	[14]	[17]	[19]	[21]	[25]	[32]
F7	[10]	[17]	[19]	[22]	[25]	[29]	[38]
F8	[10]	[17]	[18]	[22]	[25]	[28]	[32]
F9	[4]	[14]	[17]	[19]	[22]	[25]	[32]
F10	[4]	[9]	[12]	[16]	[19]	[20]	[27]
F11	[−7]	[3]	[7]	[11]	[13]	[15]	[21]
F12	[−7]	[0]	[5]	[7]	[10]	[13]	[16]

Similar comparisons with between clusters' descriptions and states individual quantile values can be made with other borderline cases brought out in Table 10. In all cases,there have been extremely cold winters, but the likelihood of them is low—the gap in degrees between 0% quantile and 10% quantile is wide.

Similar results to Fig. 8b can be followed in Irpino and Verde (2006). Despite that method considering the inner distribution of histograms, it produces similar results for macroscopic method, differences in clusters appearing with California and Arizona that are grouped similar to (a). The reason for achieving results similar to macroscopic approach is due to the fact that the length of the histogram has a large impact on their method. With proposed microscopic method, distance in every quantile value has equal impact therefore the average weather, as explained previously, is not overshadowed by extreme occurrences. The main differences between Fig. 8a and results in Irpino and Verde (2006) are with inland states that record few extremely cold years.

Table 14 Cluster quality evaluation for American States' weather

k	Calinski–Harabasz			Davies–Bouldin			RMSSTD		
	MHCC			MHCC			MHCC		
	Max	Avg	HCC	Max	Avg	HCC	Max	Avg	HCC
10	56.888	42.108	72.933	0.391	1.030	0.820	0.013	0.016	0.081
9	59.000	43.709	75.927	0.402	1.022	0.761	0.013	0.017	0.084
8	52.499	40.905	68.619	0.493	1.133	0.819	0.015	0.018	0.092
7	58.086	44.288	67.343	0.485	1.131	0.815	0.015	0.019	0.098
6	68.824	51.406	62.590	0.445	0.980	0.937	0.015	0.019	0.103
5	**79.526**	59.794	71.446	0.391	0.843	1.006	0.016	0.020	0.109
4	70.894	55.196	77.805	0.383	0.741	0.965	0.019	0.022	0.118
3	67.058	57.436	77.980	0.337	0.628	0.760	0.023	0.025	0.136
2	61.742	54.105	68.358	0.335	0.671	0.945	0.029	0.032	0.167

Cluster quality measures are calculated in Table 14. Three different indexes as previously are used. Calinski–Harabasz and RMSSTD have similar results, as with Hardwood data, indicating the best number of clusters at 5. In additionally, Calinski–Harabasz index can be compared with (Irpino and Verde 2006). They also achieved best result with 5 clusters and best maximum value of 77.65. The best value using macroscopic approach is 77.98 but this is achieved with 3 clusters—cold, mild and warm. Microscopic approach achieves best results at 5 clusters, similar to Irpino and Verde (2006), with maximum value 79.53 that is higher than previously reported.

Davies–Bouldin, on the other hand, suggests achieving best result with HCC at 3 clusters. Averaged index over all quantiles using MHCC has similar result while using the best quantile suggest using only 2 groups. DB index seems to be very sensitive to large overlap that weather dataset has.

Additionally, clusters' compactness can visually be observed in Fig. 7. The dendogram of (a) has smaller average "distance" between most of the objects than in (b). Concepts in (a) are therefore more compact. When cutting the dendrogram at 5 clusters, (a) is cut after 32% (at point 0.320) while (b) is cut at 47% (at point 0.1). The dendogram at Irpino and Verde (2006) uses Ward criterion and therefore dendograms are not comparable other than by general shape.

6 Discussion

As shown above, the main benefit of proposed algorithm is that distributional information is considered in more details allowing additional data being taken into account during conceptual clustering that is overlooked in macroscopic case. In addition, proposed approach does not suffer from computationally expensive histogram merger.

The proposed approach does not re-invent general algorithm for hierarchical conceptual clustering. The base of the algorithm is the same for both macroscopic and microscopic approach. The difference comes from how objects are compared to each

other. For macroscopic case, size of the area where probabilities can be found is considered—being it span in the case of intervals or sizes of the histograms' join and meet.

In microscopic case, differences at specific distributional points (quantiles) are measured. Individually, we could think of the distance in a single quantile point as interval's span (between minimum and maximum values of objects at that quantile) and therefore we could think of it as "macroscopic" property. For overall difference between objects, multiply quantiles in different parts of distribution are considered. We could say that we apply similar approach as span many times in different parts of distribution to get more precise result.

Due to using multiple points for comparison, mergers forming main clusters tend to happen with smaller values of dissimilarity than with macroscopic methods. For example, chaining effect as in Fig. 5b tends to occur much less with proposed algorithm than with macroscopic methods. This is due to the fact that when objects are similar, majority of the quantile values compared tend to be similar. Some differences may appear in only few individual quantiles. The overall normalized result will lower the impact of those few differences. When objects are dissimilar, all the points tend to be different in all quantile values, therefore having equal impact to normalized final dissimilarity measure. Thanks to this phenomena, it is easier to make a decision where to cut the dendrogram and decide how many clusters to keep.

There exists few other clustering methods that do considering underlying distributions. The proposed method has benefit of low required storage during algorithm execution and simple distance calculation as single point values are used.

Storage complexity of macroscopic algorithm is more complex – the whole histogram has to be stored. In general case, the histogram has many bins and those require storing three values (min, max and probability per bin). During the merger phase, new histograms are produced. To keep the exact description of merged distribution, that process produces more detailed histograms (with more bins). In worse case, if the histograms have partial overlap, actually more bins than before could be produced during merger (if original histograms had g and l bins, then merged histogram in worse case can have $g+l+1$ bins). In addition, as more mergers are done and more bins generated, the probabilities of the individual bins become very small—that is another undesired side effect of the macroscopic merger for histograms. Similar undesired side effect also occurs in Irpino and Verde (2006)'s approach using quantiles.

In case of microscopic approach, the underlying complexity of the data structure does not matter—only fixed m quantile values of every histogram is stored, no matter how many bins there are. In addition, merging does not generate new values—only minimums and maximums of quantiles are stored ($m \times 2$ values). Therefore, during clustering no new storage is required with maximum storage still being $m \times d \times n \times 2$.

As we are dealing with symbolic data, distance calculation between two objects is not doable with $O(n)$ complexity, as in classical data's case, with respect to the number of features. Instead of linear time, it becomes quadratic time $O(n^2)$ with respect of number of features and bins/quantiles. That remains true in both macroscopic and microscopic case. In macroscopic case, the most expensive operation is merger of histograms with complexity $O(n^3)$ as in Irpino and Verde (2006) and Umbleja (2017). In proposed microscopic approach, the merger is reduced to $O(n^2)$ complexity - over all

features and quantiles, simple list operation is performed to update minimum and maximum value for limited number of quantiles. In addition, now additional quantiles or bins are generated during merger (like in two previously mentioned methods)—simply current minimum and maximum value for specific quantile are stored. In addition, in macroscopic case, distance is calculated with compactness measure that depends on number of histograms already merged, $\|\omega_x\|$, similar to average histogram calculation in Irpino and Verde (2006) with complexity $O(n^3)$.

Usually, $n >> m$ and therefore number of quantiles m does not have such an huge impact of time complexity while number of objects already merged, $\|\omega_x\|$, and number of bins/quantiles in currently available algorithms tend to rise with every merger (larger the n, more mergers required). With previous approaches, the procedures (like merging histograms or calculating its compactness) are complex as they consider histogram's characteristics (like assumption of uniform distribution inside the bin and dividing it between bins when new bins/quantiles are generated). In microscopic approach, the complex nature of the underlying data can be "ignored" due to using simple fixed number quantile values. In addition, simple individual quantile values are easier to follow by human analyser than complex histograms with large number of bins. In macroscopic approach, the possibility to overcome extra bin/quantile generation can be achieved by limiting the maximum number of bins/quantiles—but that would require extra actions and modifications to current approaches.

Also, the proposed approach can be very neatly modified to match specific conditions that interests the analyser by modifying the set of quantiles used to compare objects. If only two quantiles 0% and 100% is used, the algorithm becomes similar to macroscopic approach in its analysis capacity but instead of single size, it uses two measured differences. If 0% and 100% quantiles are left out, truncated data is used. That can be beneficial for clearer picture as low probability density areas at the either side of the main data are not that important. Closer the first and last quantile values are, more of the data is left out and less amount of original distribution information is used. As normally symbolic data is aggregated from large classical data, as in case of Hardwood data, by discarding those small probability areas at the start and end of density curve removes outliers impact to aggregated data.

The number of quantiles used has an impact on the results. More quantiles used, less impact dissimilarity among one specific quantile has. At one point, adding more quantiles will not have any further influence for the results and will become counterproductive as impact of single difference is reduced to non-existent. Less quantiles, more effect every single quantile has but less details about the data are considered. Optimal choice of quantiles can be made depending on the data and the goal of the analysis but usually it would be between 4 and 10 quantiles. The amount should reflect the goal of the analysis and the volume of original data. If unaggregated data was not large, not many quantiles are needed for adequate analysis. It may be beneficial to space quantiles equally or concentrate on probability values around median.

The proposed algorithm not only forms clusters but also generates the concepts and their description. Another important benefit of proposed algoritm is that it is monotone.

7 Conclusion

In this paper, method for microscopic hierarchical conceptual clustering based on quantile values is proposed. The main contribution of proposed method is that the algorithm is simple and monotone, yet it allows considering small microscopic details in underlying distribution in distributional data. It assures that more accurate conceptual clusters are formed as a result of clustering as it was proven by applying proposed method on three commonly used datasets in symbolic data analysis. The proposed method is especially beneficial for symbolic data that is described by complex distribution functions like histograms but also has advantages to more simple distributions like intervals. Furthermore, it allows, due to usage of quantiles, comparison between different types of symbolic objects without adding complexity to the method. Algorithm can be modified, using the choice of quantiles, to match specific needs or previous knowledge about the data.

Acknowledgements The authors want to thank reviewers for their helpful comments. Kadri Umbleja's work has been supported by Japan Society for the Promotion of Science's International Research Fellow program.

8 Appendix

Implementation of algorithm in Python can be found at: https://github.com/iardacil/MHCC

References

Bertrand P, Mufti GB (2008) Stability measures for assessing a partition and its clusters: application to symbolic data sets. In: Symbolic data analysis and the SODAS software, pp 263–278

Billard L, Diday E (2006) Symbolic data analysis: conceptual statistics and data mining. Wiley, Hoboken

Brito P, De Carvalho FdA (2008) Hierarchical and pyramidal clustering. In: Symbolic data analysis and the sodas software, pp 157–180

Brito P, Ichino M (2010) Symbolic clustering based on quantile representation. In: Proceedings of COMPSTAT2010, pp 22–27

Brito P, Ichino M (2011) Conceptual clustering of symbolic data using a quantile representation: discrete and continuous approaches. In: Proceeding of theory and application of high-dimensional complex and symbolic data analysis in economics and management science, pp 22–27

de Carvalho FdA, de Souza RM (2010) Unsupervised pattern recognition models for mixed feature-type symbolic data. Pattern Recogn Lett 31(5):430–443

De Carvalho FDA, Lechevallier Y, Verde R (2008) Clustering methods in symbolic data analysis. In: Symbolic data analysis and the sodas software, pp 181–204

Diday E, Esposito F (2003) An introduction to symbolic data analysis and the sodas software. Intell Data Anal 7(6):583–601

El-Sonbaty Y, Ismail MA (1998) On-line hierarchical clustering. Pattern Recogn Lett 19(14):1285–1291

Fisher DH (1987) Knowledge acquisition via incremental conceptual clustering. Mach Learn 2(2):139–172

Goswami S, Chakrabarti A (2012) Quartile clustering: a quartile based technique for generating meaningful clusters. J Comput 4(2):48–55

Guru D, Nagendraswamy H (2006) Clustering of interval-valued symbolic patterns based on mutual similarity value and the concept of k-mutual nearest neighborhood. In: Asian conference on computer vision, Springer, Berlin, pp 234–243

Hardy A, Lallemand P (2002) Determination of the number of clusters for symbolic objects described by interval variables. In: Classification, clustering, and data analysis, Springer, Berlin, pp 311–318
Hu X (1992) Conceptual clustering and concept hierarchies in knowledge discovery. Ph.D. thesis, theses (School of Computing Science)/Simon Fraser University
Hubert L (1972) Some extensions of Johnson's hierarchical clustering algorithms. Psychometrika 37(3):261–274
Ichino M (2008) Symbolic PCA for histogram-valued data. In: Proceedings IASC, pp 5–8
Ichino M (2011) The quantile method for symbolic principal component analysis. Stat Anal Data Min: ASA Data Sci J 4(2):184–198
Ichino M, Britto P (2014) The data accumulation graph (DAQ) to visualize multi- dimensional symbolic data. In: Workshop in symbolic data analysis, Taipei, Taiwan
Ichino M, Brito P (2013) A hierarchical conceptual clustering based on the quantile method for mixed feature-type data. In: Proceedings of world statistics congress of the international statistical institute
Ichino M, Umbleja K (2018) Similarity and dissimilarity measures for mixed feature-type symbolic data. In: Studies in theoretical and applied statistics, Springer, Berlin, pp 131–144
Ichino M, Yaguchi H (1994) Generalized minkowski metrics for mixed feature-type data analysis. IEEE Trans Syst Man Cybern 24(4):698–708
Irpino A, Verde R (2006) A new wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Data science and classification, Springer, Berlin, pp 185–192
Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv (CSUR) 31(3):264–323
Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254
Jonyer I, Cook DJ, Holder LB (2001) Graph-based hierarchical conceptual clustering. J Mach Learn Res 2:19–43
Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. In: 2010 IEEE international conference on data mining, IEEE, pp 911–916
Michalski RS, Stepp RE (1983) Learning from observation: conceptual clustering. In: Machine learning, Springer, Berlin, pp 331–363
National Climatic Data Center (2014) Tables of histogram data. Climate-vegetation atlas of North America. http://www1.ncdc.noaa.gov/pub/data/cirs/drd/drd964x.tmpst.txt. Accessed 10 Aug 2015
Umbleja K (2017) Competence based learning—framework, implementation, analysis and management of learning process. Ph.D. thesis, Theses (School of Information Technologies)/Tallinn University of Technology, https://digi.lib.ttu.ee/i/?7573. Accessed 4 Oct 2018
US Geological Survey (2013) Tables of histogram data. Climate-vegetation atlas of North America. http://pubs.usgs.gov/pp/p1650-b/datatables/hgtable.xls. Accessed 24 Aug 2015
Vendramin L, Campello RJ, Hruschka ER (2010) Relative clustering validity criteria: a comparative overview. Stat Anal Data Min: ASA Data Sci J 3(4):209–235