Spectral clustering and principal component analysis as tools for variable transformation in symbolic interval-valued data ensembles

Marcin Pełka[a]

Abstract. The selection, weighting and transformation of variables are essential phases of the modelling process. Two approaches can be applied to improve a model's accuracy: the selection of variables and the transformation of variables. In symbolic data analysis, two different approaches can be adopted: principal component analysis (PCA) and spectral clustering. In all cases, we initially start with a set of symbolic variables and, after transformation, we obtain either classical variables (single numeric values) or symbolic variables that can be used in various models. The paper presents and compares PCA and spectral clustering for symbolic data when dealing with the problem of variable transformation. Artificial data with a known cluster structure was used to compare both single and ensemble clustering approaches. The results suggest that spectral clustering achieves better results for single and ensemble models.

Keywords: symbolic data analysis, ensemble learning, spectral clustering, principal component analysis

JEL: C63, C87, C90

1. Introduction

Machine learning techniques are very useful in dealing with discrimination tasks, as they are able to address various problems and are usually quite accurate. In many cases, a one-digit prediction error can be reached for

[a] Wrocław University of Economics and Business, Faculty of Economics and Finance, ul. Komandorska 118/120, 53-345 Wrocław, Poland, e-mail: marcin.pelka@ue.wroc.pl, ORCID: https://orcid.org/0000-0002-2225-5229.

different test sets (Meyer et al., 2003). Generally, it can be said that machine learning methods have reached a high level of complexity, adaptiveness, etc. and they can detect the relations and rules that occur in a dataset.

Wolpert and Macready (1997) show that the search for the best method for solving all problems associated with machine learning is useless, as no such method exists. The choice of the method should be problem-based and related to a particular classification problem. In many cases combining (aggregating) different models (methods) can prove a good solution. Models that combine the results obtained from different models is known as ensemble learning (see for example Kuncheva, 2014; Polikar, 2012; Sagi & Rokach, 2018; Zhou, 2021). Hybrid models propose a similar solution (see for example: Ardabili et al., 2019; Tsai & Chen, 2010).

In general, the main goal of cluster analysis is to obtain relatively homogeneous clusters, i.e. groups of objects that are similar when considering the variables used in cluster analysis. Usually, we would like the clusters to be isolated and cohesive (Gnanadesikan et al., 1995, p. 3). The key issue that has a major impact on the clustering process is the method of variable selection as it affects the information that will be provided to the model.

Selecting as many variables as possible seems to be inefficient and time-consuming. As we consider the variable selection for clustering, we want the final clusters to be relatively homogeneous (Gnanadesikan et al., 1995; Guyon & Elisseeff, 2003). This can be achieved by:

- selecting weights for variables (representing the variables' 'importance' in clustering);
- variable selection, where from initial n variables, we select m $(m \leq n)$. This can be seen as a special case of weighting variables, where the selected ones are assigned weight 1 and those not selected 0;

- replacing the initial n variables with new variables. This is known as variable transformation. Such new variables can have some known and desirable properties.

This paper presents and compares two methods for variable transformation for symbolic data: principal component analysis (PCA) and spectral clustering. In the empirical part of the paper, datasets with a known number of clusters are used for single and ensemble clustering methods to compare how the proposed transformation methods affect the results of clustering. All the simulations were done using the R software.

This paper is organised as follows: Section 2 introduces the main aspects of symbolic data and shows PCA and spectral clustering as techniques for symbolic variable transformation. Section 3 presents the artificial datasets and ensemble clustering for symbolic data. The results for simple and ensemble clustering that are compared according to the adjusted Rand index are described in Section 4 and Section 5 summarises the key findings.

2. Variable transformation for symbolic data

Each symbolic object can be described by different variables (see Table 1 to view some examples). These variables can be (Billard & Diday, 2006; Bock & Diday, 2000):

1. Quantitative (numerical values):

a) numerical single-valued variables;

b) numerical multi-valued variables;

c) interval-valued variables;

d) histogram variables;

2. Qualitative (categorical values):

a) categorical single-valued variables;

b) categorical multi-valued variables;

c) categorical modal variables.

Table 1. Examples of symbolic variables

Symbolic variable	Realisations	Variable type
price of a car (in EUR)	(19,000; 23,000); (20,000; 35,000); (22,000; 37,000); (32,000; 47,000)	interval-valued (non-disjoint)
engine capacity (in ccm)	(1,000; 1,200); (1,300; 1,400) (1,500; 1,800); (1,900; 2,200)	interval-valued (disjoint)
chosen car colour	{red, black, blue, yellow} {magenta, white, grey, violet}	categorical multi-valued
preferred car brand	{Toyota (0.7); Audi (0.3)} {Skoda (0.6); VW (0.3); Other (0.1)}	categorical modal
distance travelled daily [in km]	<10, 20> (0.65); <21, 30> (0.35) <10, 20> (0.40); <21, 30> (0.60)	histogram
sex of a person	{male, female}	classical (nominal)
age of the customer	20, 30, 40, 55, 24, 35, 47	classical (ratio)

Source: author's work based on: Billard and Diday (2006), Bock and Diday (2000).

Symbolic data allow us to consider the uncertainty and variability in the data, enabling the description of objects in a new, more complex way. New methods, however, are necessary to analyse this type of data.

When dealing with the issue of variable transformation for symbolic interval-valued data, two approaches can be adopted:

a) PCA for symbolic data (PCA-SDA);

b) spectral clustering for symbolic data (SPEC-SDA).

The well-known PCA for classical data involves the following steps (Hair et al., 2010; Krzanowski, 2000):

a) obtaining correlation (or covariance) matrix (**R**) for standardised data;

b) calculation of eigenvalues and eigenvectors for **R**;

c) sorting eigenvalues and eigenvectors in ascending order and selecting the first s of them. As a result, a reduced matrix is obtained;

d) multiplying the initial data matrix by the reduced eigenvalue matrix.

When dealing with interval-valued symbolic data, several approaches (algorithms) can be applied in the case of PCA-SDA.

The first proposals where the mode (the value that appears most often in a set of data values) or the average used as representative of the interval-valued data were introduced by Nagabhushan et al. (1995). Cazes et al. (1997) and Chouakria et al. (2000) proposed vertices PCA (VPCA) and centres PCA (CPCA), where the vertices of the hyperboxes or centres of the interval were used as representatives of interval-valued symbolic data.

The VPCA was improved by Lauro et al. (2000) and Douzal-Chouakria et al. (2011) by introducing a label matrix and allowing for trivial intervals and generalised weight functions.

Palumbo and Lauro (2003) proposed a midpoint and radii PCA (MRPCA) by introducing a radius to the CPCA method. D'Urso and Giordani (2004) devised a way to use least squares for MRPCA, while Gioia and Lauro (2006) introduced the application of interval algebra for all the calculations. Le-Rademacher and Billard (2012) showed how to apply covariance to extend the classical PCA. Wang et al. (2012) proposed a complete information-based PCA (CIPCA). Chen et al. (2015) defined a covariance matrix for probabilistic symbolic data and presented a new PCA based on this variance-covariance structure.

Zuccolotto (2006) suggested describing objects by estimated means of a p-dimensional variable. Oliveira et al. (2017) proposed the use of truncated versions of symbolic principal components that apply a strict subset of the original symbolic variables as a way to improve the interpretation of symbolic principal components. Ichino (2011) introduced a new quantification method for symbolic PCA. The quantile method is applied for histogram and nominal multi-value types and other types of symbolic data at the time. Su and Wu

(2024) suggested the adaptation of the symbolic PCA method for time series data.

In this paper, the CPCA, MRPCA and also methods based on the covariance matrix will be applied for dimensionality reduction in ensemble clustering methods.

In the CPCA, the $\mathbf{X}_C$ $(N \times p)$ matrix is calculated from the symbolic data matrix, where symbolic interval-valued data is replaced (substituted) by its midpoint (centre):

$$\mathbf{X}_C = \begin{bmatrix} x_{11}^c & \cdots & x_{1p}^c \\ \vdots & \ddots & \vdots \\ x_{N1}^c & \cdots & x_{Np}^c \end{bmatrix}, \tag{1}$$

where the centre is calculated as $x_{ij}^c = \frac{\underline{x}_{ij} + \overline{x}_{ij}}{2}$, with $\underline{x}_{ij}$ being the lower bound of the j-th symbolic interval-valued variable, and $\overline{x}_{ij}$ being the upper bound of the j-th symbolic interval-valued variable.

Matrix $\mathbf{X}_C$ contains the coordinates of the N hyper-rectangles. The well-known classical PCA is applied to this matrix. Then, all the vertices of each hyper-rectangle are projected in the obtained subspace and the lower-dimensional rectangles (if we extract only two principal components) are constructed with segments covering all the projections. This method assumes that the hyper-rectangle can be well-represented by its centres and then the obtained subspace optimising the projection of the centres should also be optimal for the hyper-rectangles.

The PCA for symbolic interval-valued data can be additionally done by using ranges (radii) and midpoints (centres). In this case, the covariance matrix can take the following form:

$$\mathrm{Cov}(\mathbf{X}) = \frac{1}{n}(\mathbf{X}^C)^{\mathrm{T}}(\mathbf{X}^C) + \frac{1}{n}\Delta([\mathbf{X}])^{\mathrm{T}}\Delta([\mathbf{X}]) + \frac{1}{n}[\mathbf{X}^C\Delta([\mathbf{X}]) + \Delta([\mathbf{X}])^{\mathrm{T}}\mathbf{X}^C], \quad (2)$$

where $\mathbf{X}^C$ is the matrix of the midpoints (centres) and $\Delta([\mathbf{X}])$ is the standard variance-covariance matrix calculated for single-valued data.

Two independent PCAs should be singly exploited on those two matrices which, however, do not cover the whole variance. A solution to this problem is reflected in the formula: $\mathbf{X}^C\Delta([\mathbf{X}]) + \Delta([\mathbf{X}])^{\mathrm{T}}\mathbf{X}^C$. It takes into account the residual variance simultaneously and it allows for a logical, graphical representation of data. This is a well-known PCA on the interval midpoints whose solutions are given by $\mathbf{X}^C\Sigma^{-1}\mathbf{u}_m^c = \lambda_m^c\mathbf{u}_m^c$, with λ_m^c being defined under the usual orthonormality constraints. Similarly to the PCA that is based on midpoints, the solutions are obtained for ranges $\Delta([\mathbf{X}])\Sigma^{-1}\mathbf{u}_m^r = \lambda_m^r\mathbf{u}_m^r$, with the same orthonormality constraints for λ_m^r and $\mathbf{u}_m^r$.

Palumbo and Lauro (2003) suggest maximising the convergence coefficient between the midpoints and radii proposed by Tucker:

$$f(T) = \frac{t'_l\Delta([\mathbf{X}])_l^{\mathrm{T}}\mathbf{X}^C}{\left(t'_l\Delta([\mathbf{X}])_l^{\mathrm{T}}\Delta([\mathbf{X}])_l t\right)^{1/2}\left((\mathbf{X}^C)^{\mathrm{T}}\mathbf{X}^C\right)^{1/2}}, \quad (3)$$

where: $[t_1, \ldots, t_l, \ldots, t_p]$ is the rotation matrix.

Furthermore, the PCA for symbolic interval-valued data can be done by using a covariance. In this case, the total sum of products (*SPT*) is decomposed into two components: the sum of products within (*SPW*) and the sum of products between (*SPB*), and these products are connected to the covariance:

$$nCov_{j_1,j_2} = SPT_{j_1,j_2} = SPW_{j_1,j_2} + SPB_{j_1,j_2}, \quad (4)$$

where: $CovW_{j_1,j_2} = \frac{SWW_{j_1,j_2}}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{(\overline{x}_{ij1}-\underline{x}_{ij1})(\overline{x}_{ij2}-\underline{x}_{ij2})}{12}$ and $CovB_{j_1,j_2} = \frac{SBB_{j_1,j_2}}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\overline{x}_{ij1}-\underline{x}_{ij1}}{2} - \overline{X}_{j1}\right)\left(\frac{\overline{x}_{ij2}-\underline{x}_{ij2}}{2} - \overline{X}_{j2}\right)$.

The covariance approach for PCA utilises all the information in the symbolic data. The Cov matrix is decomposed into $CovW$ and $CovB$ matrices. This allows for a deeper insight into the PCA results for traces of these matrices.

Spectral clustering is not a new clustering method, but rather a new way of preparing the dataset for other clustering methods (e.g. k-means, hierarchical clustering, etc.). Finite-sample properties of spectral clustering have been shown by Ng et al. (2002), Shi and Malik (2000).

Spectral clustering has the advantage of performing effectively in the presence of non-Gaussian clusters. Additionally, this approach is free from the drawback of the presence of local minima. The results obtained via spectral clustering in many cases outperform other well-known clustering methods (Luxburg, 2007). What is more, spectral clustering can detect clusters of different shapes, as it makes no assumptions according to the shape of clusters (Luxburg, 2007).

Spectral clustering, however, has its disadvantages. The choice of a good similarity graph is a challenging task, and, usually, it entails a fully connected graph. Spectral clustering can also be unstable under different choices of the parameters for the neighbourhood graphs. Another problem is the selection of the kernel for spectral clustering. Many different kernels can be applied and each of them can lead to different outcomes The Gaussian kernel tends to be used most often (see Karatzoglou, 2006, where the application of different kernels is presented).

Another issue in spectral clustering is the selection of a good σ parameter. This parameter should minimise the inter-cluster distances for a given number

of clusters. Karatzoglou (2006) proposed an efficient algorithm for finding the optimal σ parameter.

The spectral clustering algorithm for symbolic data involves the following steps:

1. Let **X** be the symbolic data table with n rows and m columns. Let u be the number of clusters;

2. Let $\mathbf{A} = [a_{ik}]$ be an affinity matrix for the objects. This **A** matrix can be calculated in many ways and its elements can be defined as:

$$A_{ik} = \exp(-\sigma \cdot d_{ik}) \text{ for } i \neq k, \tag{5}$$

where σ is the scaling parameter that minimises the sum of the inter-cluster distances for the given number of clusters u and d_{ik} is the distance between the i-th and k-th object;

3. Calculation of the Laplacian: $\mathbf{L} = \mathbf{D}^{1/2}\mathbf{A}\mathbf{D}^{1/2}$ (with **D** being the weight matrix with sums of each row from **A** on the diagonal);

4. Calculation of eigenvectors and eigenvalues of **L**;

5. First u eigenvectors create the **E** matrix. Each eigenvector is treated as a column of **E**, thus **E** has $n \times u$ dimensions;

6. Normalisation of **E** according to $y_{ij} = \frac{e_{ij}}{\sqrt{\sum_{j=1}^{u} e_{ij}^2}}$;

7. Finally, the **Y** matrix is the starting point for some clustering algorithms (i.e. k-means, hierarchical clustering).

The only difference between spectral clustering for classical and symbolic data lies in the applied distance measure. For details concerning distance measures for symbolic data, see Billard and Diday (2006) or Bock and Diday (2000).

3. Ensemble clustering for symbolic data

In general, ensemble learning methods are based on aggregated, combined results obtained from different models (clustering methods). These results can be seen as different points of view on the same dataset. Ensemble techniques have been applied with success in the context of supervised learning as they lead to improved accuracy and stability of algorithms (Breiman, 1996).

In ensemble clustering, we combine the results of N different models $(P_1, \dots, P_n)$ into one final clustering (aggregated clustering, ensemble clustering), i.e. P^* with k clusters (Fred & Jain, 2005).

There is a formal mathematical proof showing that in the case of ensemble learning in supervised tasks, the error reached by the ensemble is lower than any of the errors of the base models that form the ensemble (Gatnar, 2008).

Ensemble clustering can be seen as the solution to the problem with the selection of the clustering method. In this case, different clustering methods allow us to take into account 'different points of view'. Ensemble methods can prove effective when dealing with too few or too many data. If too many data occur, we can divide them into smaller, easier-to-learn partitions, and if there is a small amount of data, the same data can be used many times via bootstrapping techniques. Ensemble learning makes it also possible to deal with complex data or data too difficult to cluster. In this case, ensemble learning enables the data to be 'cut' into smaller, easier-to-learn parts, which is also known as the 'divide and conquer' approach. When dealing with many real-life problems involving decision-making, it is normal to consider information from many sources, known as information fusion (see for example Kuncheva, 2014; Zhou, 2012).

In the case of symbolic data, the following paths for ensemble clustering can be distinguished:

1. Clustering based on multiple relational matrices proposed by de Carvalho et. al. (2012). The idea is based on various distance matrices (that can be obtained from different distance measures, subsets of objects or subsets of variables). These relational matrices are used to calculate relevance weight vectors. The relevance weight vectors and distance matrices are then used to group a set of objects into final clusters;

2. Applying one of the well-known ensemble clustering methods for symbolic data:

a) proposal made by Leisch (1999), where many different clustering results are used to obtain cluster centres. Then these centres are used to obtain the final clusters. At the end, all objects are assigned to the nearest cluster. Medoids (cluster representatives) are used for symbolic data;
b) adaptation proposed by Dudoit and Fridlyand (2003), where cluster labels are permutated to get all the possible consensus clustering and all the elements of ensemble clustering;
c) Hornik's (2005) idea to minimise the distance between the set of all the possible consensus clustering elements and all elements of the ensemble clustering;

3. Applying one of the consensus functions (Fred & Jain, 2005):

a) hypergraph partitioning which assumes that clusters can be represented as edges on a graph. Their vertices correspond to the objects to be clustered. Each edge describes a set of objects belonging to the same cluster. The problem of consensus clustering is reduced to finding the minimum cut of a hypergraph;
b) the voting approach, where we permute cluster labels in such a way that the best agreement between the labels of two partitions is obtained. All the partitions from the cluster ensemble must be relabelled according to a fixed reference partition;

c) mutual information which assumes that the objective function of a clustering ensemble can be formulated as the mutual information between the empirical probability distribution of labels in the consensus partition and the labels in the ensemble. A generalised definition of mutual information is usually applied in this approach;
d) co-association-based functions where the main assumption is that objects that belong to the same cluster ('natural cluster') are co-located in the same clusters in different data partitions. The elements of the co-association matrix are defined as: $C(i,j) = \frac{n_{ij}}{N}$, where n_{ij} is the number of times that objects i and j are grouped in the same cluster (together) among all N base partitions;
e) finite mixture models where the main assumption is that the output labels are modelled as random variables drawn from a probability distribution described as a mixture of multinomial component densities. The objective of consensus clustering is formulated as a maximum likelihood estimation.

In the empirical part, Leisch's (labelled LE), Hornik's (labelled HE), Dudoid's and Fridlyand's (labelled DFE) and the co-clustering matrix (CCE) are used to obtain the final partitions (clusters) with the application of the Silhouette (Rousseeuw, 1987) clustering index to find the final number of clusters.

Although this index has some limitations, like bias toward convex or spherical clusters (see Dudek, 2020), high dimensions reduce its effectiveness (see Tomašev & Radovanović, 2016). This index is sensitive to noisy variables.

4. Single and ensemble clustering results

To compare how PCA and spectral clustering for symbolic data handle different shapes of clusters, the `cluster.Gen` function from the `clusterSim` package for the R software was used (Walesiak & Dudek, 2024). The `cluster.Gen` function allows the generation of various cluster shapes. To generate symbolic interval-valued data, the data for each model is generated twice, thanks to which sets A and B are obtained. Minimum value x_{ij}^{A}, x_{ij}^{B} is treated as the lower bound of the symbolic variable and the maximum is treated as the upper bound. The following simulation paths were used:

a) different PCA for symbolic interval-valued data were applied, then ensemble clustering methods were used (path P_1);
b) spectral clustering with different distance measures ($\sigma = 2$ in all models) was used, and the final **Y** matrix was applied for ensemble clustering (path P_2);
c) both PCA and spectral clustering were used with different initial settings (path P_3).

The following clustering methods were applied: partitioning around medoids (PAM), hierarchical-clustering (single-link), dynamic clustering for symbolic data (SClust), clustering based on the distance matrix (DClust). Both `SClust` and `DClust` are functions of the `symbolicDA` package of R.

The following datasets with known cluster structures were prepared with the application of the `cluster.Gen` function from the `clusterSim` package of the R software:

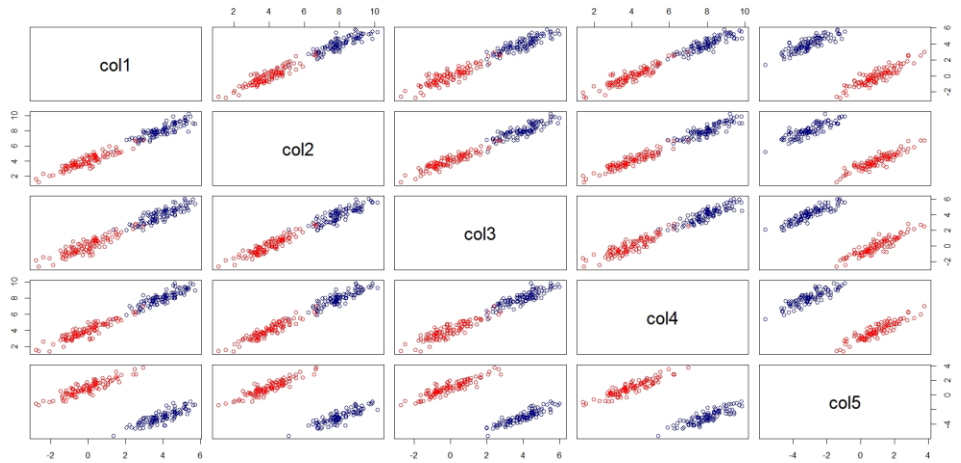
a) set I: 100 objects in two well-separated clusters (see Figure 1) in five dimensions with means $(4, 8, 4, 8, -3)$, $(0, 4, 0, 4, 1)$ and covariance matrices

$\sum_1(\sigma_{jj} = 1, \sigma_{jl} = 0.9)$, $\sum_2(\sigma_{jj} = 1, \sigma_{jl} = 0.5)$, $\sum_3(\sigma_{jj} = 1, \sigma_{jl} = -0.7)$, $\sum_4(\sigma_{jj} = 1, \sigma_{jl} = 0.84)$, $\sum_5 \begin{bmatrix} 1 & -0.45 & -0.45 & -0.45 & -0.45 \\ -0.45 & 1 & -0.56 & -0.56 & -0.56 \\ -0.45 & -0.56 & 1 & -0.58 & -0.58 \\ -0.45 & -0.56 & -0.58 & 1 & -0.74 \\ -0.45 & -0.56 & -0.58 & -0.74 & 1 \end{bmatrix}$;

b) set II: 100 objects in five not well-separated clusters in five dimensions (see Figure 2) with means (5,5,5,5,5), (−3,3,−3,3,−3), (0,0,0,0,0), (−5,−5,−5,−5,−5) and covariance matrices $\sum_1 \begin{bmatrix} 1 & -0.9 & -0.9 & -0.9 & -0.9 \\ -0.9 & 1 & -0.7 & -0.7 & -0.7 \\ -0.9 & -0.7 & 1 & -0.85 & -0.85 \\ -0.9 & -0.7 & -0.85 & 1 & -0.9 \\ -0.9 & -0.7 & -0.85 & -0.9 & 1 \end{bmatrix}$, $\sum_2(\sigma_{jj} = 1, \sigma_{jl} = 0.9)$, $\sum_3(\sigma_{jj} = 3, \sigma_{jl} = 1.5)$, $\sum_4(\sigma_{jj} = 1, \sigma_{jl} = 0)$, $\sum_5(\sigma_{jj} = 1, \sigma_{jl} = 0.2)$;

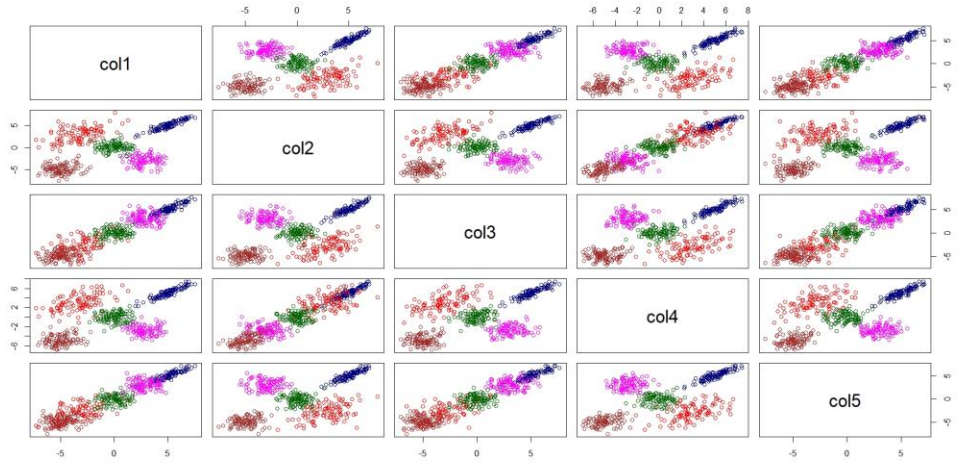
c) set III: 100 objects in three well-separated clusters in five dimensions (see Figure 3) with means $\sum_1(\sigma_{jj} = 1, \sigma_{jl} = 0)$, $\sum_2(\sigma_{jj} = 1, \sigma_{jl} = -0.9)$, $\sum_3(\sigma_{jj} = 1, \sigma_{jl} = 0.9)$, $\sum_4(\sigma_{jj} = 3, \sigma_{jl} = 1)$, $\sum_5(\sigma_{jj} = 1, \sigma_{jl} = -0.5)$.

Figure 1. Two well-separated clusters in five dimensions (set I)



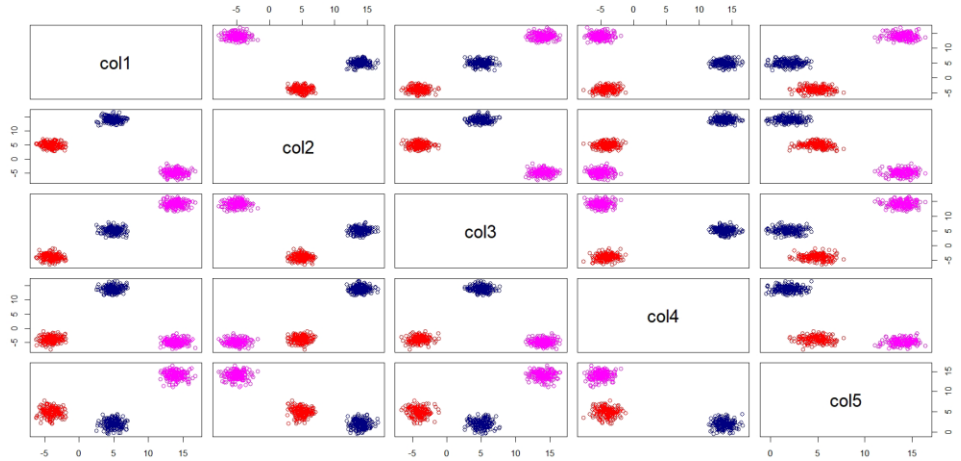
Source: author's work based on the R software.

Figure 2. Five not well-separated clusters in five dimensions (set II)



Source: author's work based on the R software.

Figure 3. Three well-separated clusters in five dimensions (set III)



Source: author's work based on the R software.

To compare how PCA and spectral clustering perform for symbolic data, 50 simulations were done and the average adjusted Rand index (Rand, 1971) was calculated. The average adjusted Rand index values are shown in Table 2 for single clustering methods and in Table 3 for ensemble clustering results.

Table 2. Simulation results for single models: all paths and datasets

Single model	Dataset	Adjusted Rand index value	Calculation time
pam (cPCA)	I	0.9984	2.95938 mins
	II	0.9862	3.68677 mins
	III	1	3.50588 mins
pam (mrPCA)	I	1	2.94649 mins
	II	0.9982	3.92382 mins
	III	0.7349	3.67878 mins
pam (covPCA)	I	0.9943	4.12835 mins
	II	1	2.66233 mins
	III	1	57.90654 secs
pam (specl & H)	I	0.9959	2.96048 mins
	II	0.9801	2.23560 mins
	III	0.9978	4.96317 mins
pam (specl & U_2)	I	1	2.90862 mins
	II	0.9993	3.01923 mins
	III	1	4.97362 mins
pam (specl & SO_1)	I	0.9424	2.60051 mins
	II	0.9720	7.33325 mins
	III	1	5.86112 mins
DClust (cPCA)	I	1	3.58567 mins
	II	0.5912	10.95904 mins
	III	1	1.93726 mins
DClust (mrPCA)	I	0.7723	1.72763 mins
	II	0.3434	10.71638 mins
	III	1	3.84637 mins
DClust (covPCA)	I	0.9882	1.72323 mins
	II	0.5864	10.70673 mins
	III	0.5567	3.89015 mins
DClust (specl & H)	I	1	2.54698 mins
	II	0.8065	16.12656 mins
	III	0.8899	11.30388 mins
DClust (specl & U_2)	I	0.9899	13.63857 mins
	II	0.7422	16.00192 mins
	III	0.7597	5.76498 mins
DClust (specl & SO_1)	I	0.9616	2.61785 mins
	II	0.5486	15.94209 mins
	III	1	5.74380 mins

Note. pam – partition around medoids, DClust – dynamic clustering based on the distance matrix, cPCA – centres PCA, mrPCA – midpoints and radii PCA, covPCA – covariance-based PCA, specl – spectral clustering, H – Hausdorff distance, U_2 – Ichino-Yaguchi distance, SO_1 – de Carvalho distance.
Source: author's work based on the R software.

Table 3. Simulation results for ensemble models: all paths and datasets

Ensemble model	Dataset	Adjusted Rand index value		
		Path P_1	Path P_2	Path P_3
LE	I	0.8728	0.8989	0.9014
	II	0.6533	0.7623	0.9765
	III	0.8672	0.9123	0.9826

HE	I	0.9836	1	1
	II	0.9123	1	1
	III	0.9635	1	1
DFE	I	0.9927	1	1
	II	0.9563	0.9991	0.9831
	III	0.9972	1	1
CCE	I	0.9873	1	1
	II	0.9654	0.9864	1
	III	0.9862	1	1

Note. LE – Leich's ensemble, HE – Hornik's ensemble, DFE – Dudoit and Fridlyand's ensemble, CCE – co-clustering matrix ensemble.
Source: author's work based on the R software.

Dataset II (five not well-separated clusters in five dimensions) was the most challenging to cluster for all of the methods. However, the classical pam method combined with either the PCA or the spectral approach for symbolic data outperformed the DClust method designed for symbolic data.

When we look at the ensemble results (Table 3), we can see that Hornik's ensemble model, as well as Dudoit's and Fridlyand's, ensemble models achieved the highest average values of the adjusted Rand index across all model types. Similarly to the single model results, it was the most challenging to detect clusters by the ensemble models in Dataset II. Nevertheless, Hornik's, Dudoit's and Fridlyand's ensemble models perform most efficiently.

5. Conclusions

Two different approaches for symbolic data transformation have been shown in the paper for ensemble learning with this type of data. The first approach uses the well-known PCA applied for symbolic data, the second one utilises spectral clustering. Additionally, a combination of PCA and the spectral approach was used in the ensembles.

However, PCA for symbolic data is limited to symbolic interval-valued data only, while spectral clustering for symbolic data can handle various symbolic data types, as it requires only an appropriate distance measure for symbolic

data. Notably, there are many different symbolic distance measures suitable for various symbolic variable types.

Ensemble clustering for symbolic data enables the integration of different clustering results (‘points of view’) to achieve a single, improved and more stable clustering outcome. In ensemble clustering, the key steps such as variable selection, variable weighting and variable transformation remain critical, as in the case of a single clustering method.

The results indicate that complex datasets (e.g. those with intricate cluster structures, outliers or noisy variables) are challenging for single symbolic clustering methods based on PCA. However, PCA combined with spectral clustering, as well as spectral clustering alone performs most effectively with such datasets (as measured by the Adjusted Rand Index). Similar trends are observed with ensemble clustering methods for symbolic data. Specifically, symbolic ensemble clustering techniques such as Hornik’s, Dudoit’s and Fridlyand’s methods generally outperform a co-clustering (co-occurrence) matrix and Leisch’s ensemble methods when dealing with complex data structures.

References

Ardabili, S., Mosavi, A., & Várkonyi-Kóczy, A. R. (2019). Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods. In A. R. Várkonyi-Kóczy (Ed.), *Engineering for Sustainable Future. Selected papers of the 18th International Conference on Global Research and Education Inter-Academia – 2019* (pp. 215–227). Springer. https://doi.org/10.1007/978-3-030-36841-8_21.

Bock, H.-H., & Diday, E. (2000). *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer. https://doi.org/10.1007/978-3-642-57155-8.

Billard, L., & Diday, E. (2006). *Symbolic Data Analysis. Conceptual Statistics and Data Mining*. Wiley. https://doi.org/10.1002/9780470090183.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655.

de Carvalho, F. A. T., Lechevallier, Y., & de Melo, F. M. (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, *45*(1), 447–464. https://doi.org/10.1016/j.patcog.2011.05.016.

Cazes, P., Chouakria, A., Diday, E., & Schektrman, Y. (1997). Extensions de l'Analyse en Composantes Principales à des données de type intervalle. *Revue de Statistique Appliquée*, *45*(3), 5–24. http://eudml.org/doc/106421.

Chen, M., Wang, H., & Quin, Z. (2015). Principal component analysis for probabilistic symbolic data: a more generic and accurate algorithm. *Advances in Data Analysis and Classification*, *9*, 59–79. https://doi.org/10.1007/s11634-014-0178-2.

Chouakria, A., Diday, E., & Cazes, P. (2000). Vertices Principal Components Analysis With an Improved Factorial Representation. In A. Rizzi, M. Vichi & H.-H. Bock (Eds.), *Advances in Data Science and Classication* (pp. 397–402). Springer. https://doi.org/10.1007/978-3-642-72253-0_54.

Dudek, A. (2020). Silhouette Index as Clustering Evaluation Tool. In K. Jajuga, J. Batóg, M. Walesiak (Eds.), *Classification and Data Analysis. SKAD 2019. Studies in Classification, Data Analysis, and Knowledge Organization.* Springer. https://doi.org/10.1007/978-3-030-52348-0_2.

Douzal-Chouakria, A., Billard, L., & Diday, E. (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining*, *4*(2), 229–246. https://doi.org/10.1002/sam.10118.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, *19*(9), 1090–1099. https://doi.org/10.1093/bioinformatics/btg038.

D'Urso, P., & Giordani, P. (2004). A least squares approach to principal component analysis for interval valued data. *Chemometrics and Intelligent Laboratory Systems*, *70*(2), 179–192. https://doi.org/10.1016/j.chemolab.2003.11.005.

Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clustering using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(6), 835–850. https://doi.org/10.1109/TPAMI.2005.113.

Gatnar, E. (2008). *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*. Wydawnictwo Naukowe PWN.

Gioia, F., & Lauro, C. N. (2006). Principal component analysis on interval data. *Computational Statistics*, *21*(2), 343–363. https://doi.org/10.1007/s00180-006-0267-6.

Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, *12*, 113–136. https://doi.org/10.1007/BF01202271.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182. https://doi.org/10.1162/153244303322753616.

Hair, J. E., Black, W. C., Babin, J. B., & Anderson, R. E. (2010). *Multivariate Data Analysis* (17th edition). Prentice Hall.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software*, *14*(12), 65–72. https://doi.org/10.18637/jss.v014.i12.

Ichino, M. (2011). The quantile method for symbolic principal component analysis. *Statistical Analysis and Data Mining*, *4*(2), 184–198. https://doi.org/10.1002/sam.10111.

Karatzoglou, A. (2006). *Kernel Methods. Software, Algorithms and Applications* [Doctoral dissertation, Vienna University of Technology]. https://resolver.obvsg.at/urn:nbn:at:at-ubtuw:1-14467.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis. A User's Perspective*. Oxford University Press. https://doi.org/10.1093/oso/9780198507086.001.0001.

Kuncheva, L. I. (2014). *Combining Pattern Classifiers. Methods and Algorithms* (2nd edition). John Wiley and Sons.

Lauro, N. C., Verde, R., & Palumbo, F. (2000). Factorial Methods with Cohesion Constraints on Symbolic Objects. In H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, M. Schader (Eds.), *Data Analysis, Classification, and Related Methods* (pp. 381–386). Springer. https://doi.org/10.1007/978-3-642-59789-3_60.

Leisch, F. (1999). *Bagged clustering* (SFB Working Papers No. 51). https://doi.org/10.57938/9b129f95-b53b-44ce-a129-5b7a1168d832.

Le-Rademacher, J., & Billard, L. (2012). Symbolic covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*, *21*(2), 413–432. https://doi.org/10.1080/10618600.2012.679895.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Meyer, D., Leisch, F., & Hornik, K. (2003), The support vector machine under test. *Neurocomputing*, *55*(1–2), 169–186. https://doi.org/10.1016/S0925-2312(03)00431-4.

Nagabhushan, P., Chidananda Gowda, K., & Diday, E. (1995). Dimensionality reduction of symbolic data. *Pattern Recognition Letters*, *6*(2), 219–223. https://doi.org/10.1016/0167-8655(94)00085-H.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On Spectral Clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems* (pp. 849–856). MIT Press.

Oliveira, M. R., Vilela, M., Pacheco, A., Valadas, R., & Salvador, P. (2017). Extracting Information from Interval Data Using Symbolic Principal Component Analysis. *Austrian Journal of Statistics*, *46*(3–4), 79–87. https://doi.org/10.17713/ajs.v46i3-4.673.

Palumbo, F., & Lauro, C. N. (2003). A PCA for interval-valued data based on midpoints and radii. In H. Yani, A. Okada, K. Shigemasu, Y. Kano, J. J. Meulman (Eds.), *New Developments in Psychometrics* (pp. 641–648). Springer. https://doi.org/10.1007/978-4-431-66996-8 74.

Polikar, R. (2012). Ensemble learning. In C. Zhang, & Y. Ma (Eds.), *Ensemble Machine Learning* (pp. 1–34). Springer. https://doi.org/10.1007/978-1-4419-9326-7_1.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/2010.1016/0377-0427(87)90125-7.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/2010.1002/widm.1249.

Shi, J., & Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905. https://doi.org/10.1109/34.868688.

Su, E. C.-Y., & Wu, H.-M. (2024). Dimension reduction and visualization of multiple time series data: a symbolic data analysis approach. *Computational Statistics*, *39*(4), 1937–1969. https://doi.org/10.1007/s00180-023-01440-7.

Tomašev, N., & Radovanović, M. (2016). Clustering Evaluation in High-Dimensional Data. In M. Celebi & K. Aydin (Eds.), *Unsupervised Learning Algorithms* (pp. 71–107). Springer. https://doi.org/10.1007/978-3-319-24211-8_4.

Tsai, C. F., & Chen, M. L. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, *10*(2), 374–380. https://doi.org/10.1016/j.asoc.2009.08.003.

Walesiak, M., & Dudek, A. (2024). *Package 'clusterSim': Searching for Optimal Clustering Procedure for a Data Set*. R package version 0.51-5. https://cran.r-project.org/web/packages/clusterSim/clusterSim.pdf.

Wang, H., Guan, R., & Wu, J. (2012). CIPCA: complete-information-based principal component analysis for interval-valued data. *Neurocomputing*, *86*, 158–169. https://doi.org/10.1016/j.neucom.2012.01.018.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893.

Zhou, Z. H. (2012). *Ensemble Methods. Foundations and Algorithms*. CRC Press. https://tjzhifei.github.io/links/EMFA.pdf.

Zhou, Z. H. (2021). *Ensemble Learning*. Springer. https://doi.org/10.1007/978-981-15-1967-3.

Zuccolotto, P. (2006). Principal components of sample estimates: an approach through symbolic data analysis. *Statistical Methods & Applications*, *16*(2), 173–192. https://doi.org/10.1007/s10260-006-0024-6.