Handling reliability big data: a similarity-based approach for clustering a large fleet of assets

Francesco Cannarile[1,2], Michele Compare[1,2], Francesco Di Maio[1*], Enrico Zio[1,2,3]

[1]*Energy Department, Politecnico di Milano, Via la Masa 34, 20156 Milano, Italy*

[2] *Aramis Srl, Via pergolesi 5, Milano, Italy*

[3] *Chair on System Science and the Energetic Challenge, Fondation EDF, Centrale Paris and Supelec, Paris, France*

Abstract

Collecting reliability field data from a very large fleet of assets poses the problem of how to effectively exploit such big data to optimize the asset maintenance strategies. To address this issue, in this work we propose a clustering algorithm based on the similarity of the asset failure behaviors: the identification of assets with similar reliability distributions enables addressing the maintenance optimization problem of all the assets belonging to the same cluster. To develop this approach, the numerous assets are first grouped into populations, based on their covariates (e.g., the working conditions and the location). Then, for each population the reliability distribution is inferred from the corresponding failure data, and its similarity with the corresponding distributions of the other populations is evaluated by calculating the Symmetric Kullback-Leibler Divergence (SKLD). The obtained similarity values are fed in input to a spectral clustering algorithm, which finds the clusters of assets that will be treated as a whole by the maintenance decision maker.The proposed approach is applied to a real case study concerning a set of more than 30000 switch point machines.

Key Words*:* Reliability, Big Data, Unsupervised Clustering, Switch Point Machines

1. INTRODUCTION

The optimization of the number of maintenance strategies based on reliability field data becomes a challenging issue when the number of managed assets is very large (e.g., millions). The proper use and exploitation of the associated reliability (big) data calls for advanced data mining techniques (Meeker & Hong, 2014). In this context, this work proposes a clustering algorithm for grouping assets based on their reliability distributions: clusters of assets with similar failure behavior will, then, undergo the same preventive maintenance optimal schedule. Indeed, this approach opens to the possibility of optimizing the maintenance strategy specifically for all assets belonging to a same cluster.

The proposed approach has been applied to reliability data collected for devices of the Italian railway system. Before the study presented in this paper, the grouping of the assets (for maintenance strategies planning) was only based on some technical information (such as rail type, switch point machine model, etc.) and/or geographic localization, rather than accounting for their common reliability features, that is, indeed the key novelty of the here proposed approach. Although the case study is derived from a real industrial application, due to the non-disclosure agreement with the industrial partner, the data shown in this paper have been opportunely re-scaled and modified.

The remainder of the paper is organized as follows: Section 2 states the problem. Section 3 details the methodology to cluster assets based on their reliability distributions. Section 4 describes the application of the Symmetric Kullback-Leibler Divergence (SKLD) to assess the similarity between reliability distributions. In Section 5, the spectral clustering algorithm is presented. Section 6 presents the case study concerning assets of the Italian railways. Finally, in Section 7 some conclusions are drawn.

2. PROBLEM STATEMENT

Our objective is to develop a clustering algorithm capable of grouping A different assets into C^* clusters, $C^* \leq A$, based on their reliability behaviors so that the decision maker would reduce the number of maintenance decisions to take from A to C^*. We suppose that for the a^{th} asset, $a \in \{1,\ldots,A\}$, the following pieces of information are available:

1. the values of K categorical variables (covariates) $(X_1,...,X_K)$ containing technical information about the asset. These values are collected into the vector $\boldsymbol{x}_a = \left(x_a^1,\ldots,x_a^K\right)$;
2. A collection of n_a independent field observations registered into the vector $\mathbf{D}_a = \left(\boldsymbol{y}_a, \boldsymbol{\delta}_a\right)$, where $\boldsymbol{y}_a = \left(y_a^1,\ldots,y_a^{n_a}\right)$, $y_a^b \in \mathbb{R}_0^+$, $b = 1,...,n_a$ and $\boldsymbol{\delta}_a = \left(\delta_a^1,\ldots,\delta_a^{n_a}\right)$ $\delta_a^b \in \{0,1\}, b = 1,...,n_a$; in other words y_a^b is a current failure time if δ_a^b is equal to 1, or a right-censored observation if δ_a^b is equal to 0, $b \in \{1,\ldots,n_a\}$.

Finally, we assume that a perfect maintenance is performed upon asset failure and, then, the asset can be considered "as good as new" after repair. This assumption allows us considering failure times within dataset $\boldsymbol{O}_a$ as statistically independent.

3. METHODOLOGY SNAPSHOT

The methodology proposed in this work is based on the following steps:

1 Based on the knowledge of experts, identify the subset $(\tilde{X}_1,\ldots,\tilde{X}_{\tilde{K}})$ of covariates $(X_1,\ldots,X_K)$, $\tilde{K}<K$, which allows partitioning the assets in $N<A$ populations, corresponding to different combinations of covariates $(\tilde{X}_1,\ldots,\tilde{X}_{\tilde{K}})$. Then, each statistical population $i=1,\ldots,N$ is assigned a time to failure dataset $\mathbf{O}_i=\left(\mathbf{D}_{i_1},\ldots,\mathbf{D}_{i_{n_i}}\right)$, where $\left\{i_1,\ldots,i_{n_i}\right\}$ are indexes referring to the assets belonging to the i^{th} population. Notice that the identification of these populations differs from clustering analysis, which, indeed, aims at grouping these statistical populations based on their reliability distributions.

In particular, the reliability distribution of the i^{th} population is assumed to be a Weibull distribution of scale parameter α_i and shape parameter β_i, $i=1,\ldots N$. The probability density function is given by

$$f_i\left(y|\alpha_i,\beta_i\right)=\frac{\beta_i}{\alpha_i}\left(\frac{y}{\alpha_i}\right)^{\beta_i-1}e^{-\left(\frac{y}{\alpha_i}\right)^{\beta_i}}\quad y>0,\ \alpha_i>0,\ \beta_i>0 \tag{1}$$

whereas the corresponding reliability function and hazard rates are, respectively:

$$R_i\left(y|\alpha_i,\beta_i\right)=e^{-\left(\frac{y}{\alpha_i}\right)^{\beta_i}}\quad y>0,\ \alpha_i>0,\beta_i>0 \tag{2}$$

and

$$h_i\left(y|\alpha_i,\beta_i\right)=\frac{\beta_i}{\alpha_i}\left(\frac{y}{\alpha_i}\right)^{\beta_i-1}\quad y>0,\ \alpha_i>0,\ \beta_i>0 \tag{3}$$

2 Apply the Maximum Likelihood Estimation (MLE) technique to each population i to estimate the parameters (α_i,β_i) of Eqs. (1), (2) and (3). In this respect, in this work we assume that MLE exists for every $i=1,\ldots,N$. In many real applications, this assumption may not be valid. In these cases, one can use either Bayesian statistical inference techniques (Cannarile et al., 2016), or perform quantization on the $\tilde{K}$ covariates to reduce the number of possible populations, each one consisting which of a larger number of assets.

3 Quantify the similarity between all pairs of statistical populations from the reliability perspective. This task is achieved quantifying how similar are the reliability distributions (either Eq. (1) or Eq. (2)) of each pair of statistical populations i and j, where $i,j=1,\ldots,N$.. The SKLD is here adopted to compute the similarity w_{ij} between densities f_i and f_j representative of the reliability

distributions of statistical populations i and j, respectively, This point of the methodology is detailed in Section 4.

4 The similarity matrix W, whose entries are given by similarities w_{ij}, is given in input to the Spectral Clustering Algorithm (SCA). This point of the methodology is detailed in Section 5.

5 Infer the best number of clusters C^* quantifying a compromise between the silhouette (Rousseuw, 1987) and Davies-Bauldin coefficients (Davies & Bauldin, 1979), as described in Section 6.

6 Create the failure time datasets $\mathbf{O}_p = \left(\mathbf{O}_{p_1}, \ldots, \mathbf{O}_{p_{n_p}} \right)$, where $\left\{ p_1, \ldots, p_{n_p} \right\}$ are the indexes referring to the statistical population assigned to cluster $p \in \left\{ 1, \ldots, C^* \right\}$. Again, each cluster is assumed to be Weibull distributed with scale parameter α_p and shape parameter β_p.

7 Apply the MLE technique to estimate parameters (α_p, β_p) for each cluster.

4. SIMILARITY BETWEEN PROBABILITY DISTRIBUTIONS

In this work, we need a similarity measure (Pollard, 2002) that takes large values for probability distributions describing similar reliability behaviours, and small values for those modelling diverse reliability behaviors.

However, when dealing with probability distributions, similarity measures such as the Euclidean distance, Mahalanobis distance, etc. cannot be exploited, since they apply to finite-dimensional objects, whereas probability distributions are infinite. For this reason, we use the Kullback-Leibler Divergence (KLD) dissimilarity measure (Kullback & Leibler, 1951).

Throughout this paper, let Ω denote the sample space, F a σ-algebra on Ω, and P the set of probability measures on the measurable space (Ω, F). Let μ_i and μ_j be two elements of P, and f_i and f_j their corresponding probability density functions with respect to a dominating measure ρ. Then, the KLD dissimilarity of μ_i from μ_j, denoted with $d_{KL}(\mu_i \| \mu_j)$, is defined as:

$$d_{KL}(\mu_i \| \mu_j) = d_{KL}(f_i \| f_j) = \int_{\Omega} f_i \ln \frac{f_i}{f_j} d\rho \tag{4}$$

Note that, in general, $d_{KL}(\mu_i \| \mu_j) \neq d_{KL}(\mu_j \| \mu_i)$.

Otherwise, if we define the SKLD between μ_i and μ_j as:

$$d_{KL}^{sym}\left(\mu_i, \mu_j\right) = \frac{1}{2}\left(d_{KL}(\mu_i \| \mu_j) + d_{KL}(\mu_j \| \mu_i) \right) \tag{5}$$

then d_{KL}^{sym} is a dissimilarity measure (being symmetric). We can, therefore, define the similarity corresponding to the SKLD (Gower, 1985) (Gower, 1986) as in Eq. (6):

$$w_{ij} = \frac{1}{1+d_{KL}^{sym}} \tag{6}$$

This measure is used to compute the similarity between the reliability distributions f_i and f_j of two different populations of components. In this respect, notice that the densities f_i and f_j in Eqs. (4) and (5) are assumed Weibull distributions in our case study. This makes the computation of w_{ij} not straightforward. Nonetheless, we can exploit the results provided in (Bauckhage, 2013) to efficiently compute the KLD divergence in Eq. (4) between two Weibull densities $f_i\left(y\,|\,\alpha_i,\beta_i\right)$ and $f_j\left(y\,|\,\alpha_j,\beta_j\right)$ as in Eq. (7):

$$d_{KL}(f_i\,||f_j) = \log\left(\frac{\beta_i}{\alpha_i^{\beta_i}}\right) - \log\left(\frac{\beta_j}{\alpha_j^{\beta_j}}\right) - \left(\beta_i-\beta_j\right)\left(\log\left(\alpha_i\right) - \frac{\gamma}{\beta_i}\right) + \left(\frac{\alpha_i}{\alpha_j}\right)^{\beta_j}\Gamma\left(\frac{\beta_j}{\beta_i}+1\right) - 1 \tag{7}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant, whereas Γ is the gamma function defined as follows:

$$\Gamma\left(z\right) = \int_0^{+\infty} t^{z-1}e^{-t}dt \;\; z \geq 0 \tag{8}$$

5. SPECTRAL CLUSTERING

The computation of the similarity between all pairs of f_i and f_j to be clustered by using similarity in Eq. (6) originates the similarity matrix W of size $\left(N,N\right)$, whose generic element w_{ij} represents the similarity between the statistical populations i and j (and thus, the diagonal elements w_{ii} are set to 1 and the matrix is symmetric $w_{ij} = w_{ji}$). From matrix W, a similarity graph $G = \left(V,E\right)$ is constructed, where each vertex v_i represents the i-th group and the weight associated to the edge e_{ij} connecting the two vertices i and j is the similarity value w_{ij} (von Luxburg, 2007). In this view, the original problem of identifying families of similar statistical populations is re-formulated in that of finding the partition of the similarity graph such that the edges connecting elements of different groups have the smallest weights, whereas the edges connecting elements within the same group have the largest weights (Alpert et al., 1999).

In details, the spectral clustering algorithm is based on the following steps (Baraldi et al., 2013):

Step1: normalized Graph Laplacian Matrix

Compute:

- the degree matrix D which is a diagonal matrix with diagonal entries $d_1,\ldots,d_N$ defined by:

$$d_i = \sum_{j=1}^{N} w_{ij}, \quad i = 1,\ldots,N \tag{9}$$

- the normalized graph Laplacian matrix

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{10}$$

where $L = D - W$, and I is the identity matrix of size (N, N).

Step2: feature extraction

The relevant information on the structure of the matrix W is obtained by considering the eigenvectors $u_1,\ldots,u_C$ associated to the C smallest eigenvalues $\lambda_1,\ldots,\lambda_C$ of its laplacian matrix L_{sym}, where C is the desired number of clusters. The square matrix W is transformed into a reduced matrix U of size (N, C), in which the C columns of U are the eigenvectors $u_1,\ldots,u_C$. Thus, the i-th object is captured in the C-dimensional vector u_i corresponding to the i^{th} row of the matrix U. A matrix T is formed from U by normalizing its row (von Luxburg, 2007):

$$t_{ic} = \frac{u_{ic}}{\left(\sum_{c=1}^{C} {u_{ic}}^2\right)^{0.5}}, i = 1,\ldots,N, c = 1,\ldots,C \tag{11}$$

It has been shown that this change of representation enhances the cluster properties in the data, so that clusters can be more easily identified (von Luxburg, 2007).

Step3: Unsupervised clustering

In this work, we resort to the K-means (Hartigan, 1975) algorithm to partition the data into C clusters. Details on this clustering method can be found in Appendix.

6. CASE STUDY

A case study concerning devices of the railway infrastructure is here discussed. The available dataset consists of $A = 32285$ different assets for which the values of $K = 12$ categorical variables $(X_1,\ldots, X_K)$ are provided (for confidentiality, details are not given here). Among these covariates, a subset

of $\tilde{K}=5$ categorical variables $(\tilde{X}_1,\ldots,\tilde{X}_{\tilde{K}})$ has been selected by experts. Based on their values, $N=374$ populations of components have been identified, with corresponding failure time datasets $\boldsymbol{O}_i$, $i\in\{1,\ldots,374\}$. Then, the estimates of the Weibull parameters (α_i,β_i) for all 374 populations have been obtained by resorting to MLE method. In Figure 1, the estimated values of the scale parameters (abscissas, in logarithmic scale) and shape parameters (ordinates) are shown.

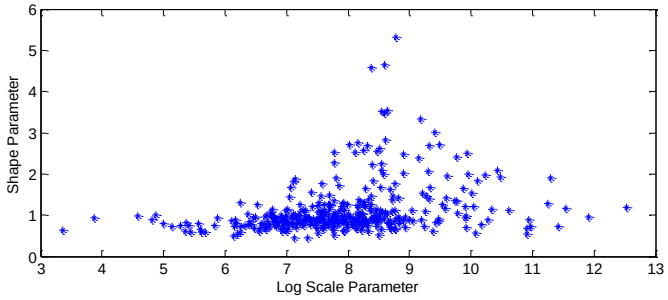

Figure 1: Estimated Weibull parameters for each statistical population

The similarity matrix W has been obtained by computing the similarity measure w_{ij} of Eq (6) between all possible 374 pairs of statistical populations. To assess the most appropriate number of clusters, we have resorted to the silhouette and Davies-Bauldin coefficients. The former measures how similar that population is to the populations in its own cluster with respect to the populations in other clusters, and ranges from -1 to +1, where values close to one indicate a good clustering. The latter is based on a ratio of within-cluster and between-cluster distances, and therefore, the smaller the Davies-Bouldin index value, the better the clustering. Figures 2 and 3, respectively, show the values of these coefficients in correspondence to the number of clusters varying from 2 to 10.

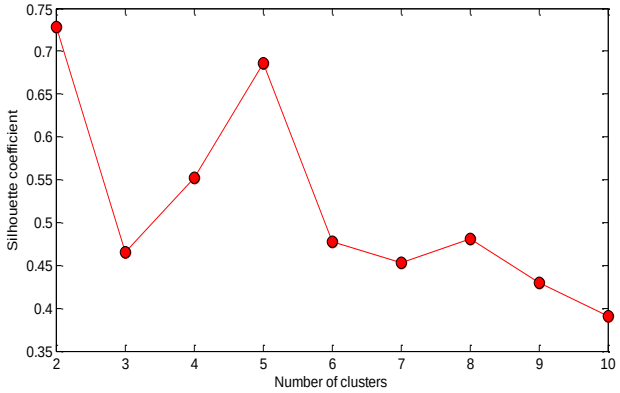

Figure 2: Silhouette coefficient increasing the number of clusters from 2 to 10.

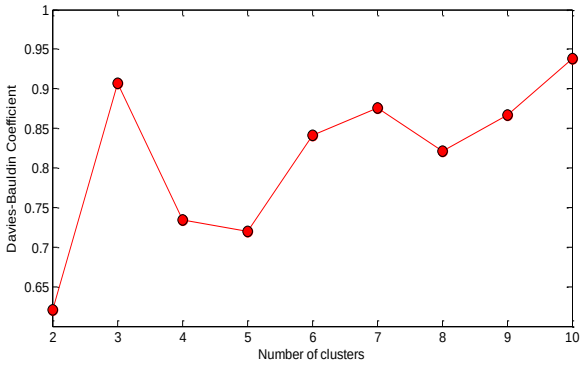

Figure 3: Davies-Bauldin coefficient increasing the number of clusters from 2 to 10.

Excluding the case in which only two clusters are identified, from both Figures 2 and 3 we can conclude that the best compromise solution according to the two coefficients is given when $C^* = 5$. Figure 4 shows, for each statistical population, in abscissa the log scale parameter ($\log(\alpha)$) and in ordinate the respective value of the shape parameter (β). From this Figure, it emerges that the 5 clusters (depicted by different markers) divide the semi-plane $(\log(\alpha)), \beta)$ in 5 pairwise disjoint regions and, therefore, we can conclude that these 5 clusters really identify different reliability behaviours.

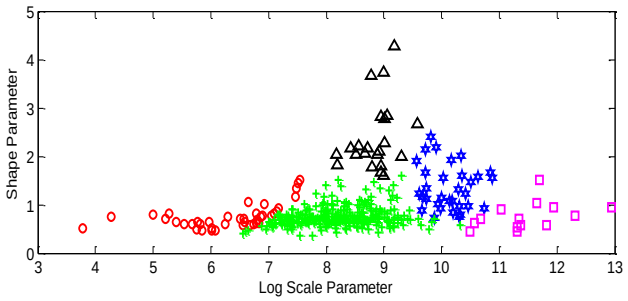

Figure 4: Values of the (log) scale parameters (abscissa axis) and shape parameters (ordinate axis) (different markers correspond to the different clusters).

After these clusters have been identified, we can estimate the common reliabity distribution of all assets belonging to the same clusters. We have still assumed that the reliability behaviour of each clusters is described by a Weibull probability distribution, by reason of the flexibility of this distribution. In Table 1, the MLE values of the scale parameters and shape parameters are reported for each cluster. From this, one can conclude that:

1. There are three clusters (circles, crosses and squares) for which the estimated values of the shape parameters are similar to each other ($\beta < 1$), whereas the estimated values of the scale parameters are very different. For these clusters, the hazard rate is a decreasing function of time.

2. There are two clusters (stars and triangles) with shape parameters assuming values larger than one, and with very different values of the estimated scale parameters. For these clusters, the failure rate is an increasing function of time.

Scale Parameter α_p	Shape Parameter β_p	Cluster marker
0.0039 e+05	0.7268	circle
0.3376e+05	1.3475	star
0.0251e+05	0.8263	cross
0.0683e+05	2.4293	triangle
1.8325e+05	0.8783	square

Table1: MLEs of scale and shape parameters of each cluster.

Finally, in Figure 5, the reliability functions relative to the $C^* = 5$ identified clusters are shown. This information enables the scheduling of only 5 preventive maintenance strategies, which are applied to all the assets belonging to the corresponding clusters.

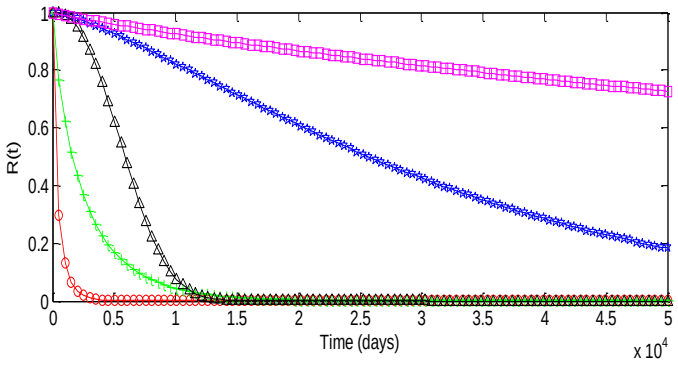

Figure 5: Reliability function $R(t)$ relative to the 5 clusters (different markers correspond to the different clusters)

7. CONCLUSIONS

In this work, we have presented a similarity-based approach for managing reliability big data to optimize maintenance strategies on large fleets of assets. Our methodology is based on, firstly, grouping the assets in statistical populations according to their technical properties, then, clustering them based on the similarity of their reliability functions. To quantify the similarity between reliability functions, the SKLD has been exploited. The proposed methodology has been successfully

applied to a case study concerning assets of the railway system, with more than 30000 assets grouped in 5 clusters. This way, only 5 preventive maintenance strategies need to be optimally scheduled, and applied to all the assets belonging to the corresponding clusters.

8. REFERENCES

Alpert, C., Kahng, A., Yao, S. 1999. Spectral partitioning: the more eigenvectors, the better. *Discrete Applied Math* 90: 3-26.

Baraldi, P., Di Maio, F., Zio, E. 2013. Unsupervised Clustering for Fault Diagnosis in Nuclear Power Plant Components. *International Journal of Computational Intelligence Systems*,6 (4): 764-777.

Baraldi, P., Di Maio, F., Rigamonti, M., Zio, E., Seraoui, R. (2015). Clustering for unsupervised fault diagnosis in nuclear turbine shut-down transients. *Mechanical Systems and Signal Processing* 58: 160-178.

Bauckhage, C. 2013. Computing the Kullback-Leibler Divergence between two Weibull Distributions. *arXiv:1310.3713 [cs.IT]*.

Cannarile, F., Compare, M., Mattafirri, S., Carlevaro, F., Zio, E. 2016. Comparison of Weibayes and Markov Chain Monte Carlo methods for the reliability analysis of turbine nozzle components with right censored data only. *Safety and Reliability: Methodology and Applications - Proceedings of the European Safety and Reliability Conference, ESREL 2015*.

Cavanaugh, J.E. 1999.A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics and Probability Letters* 42 (4): 333-343.

Christensen, R., Johnson, W., Branscum, A., Hanson, T. 2010. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. CRC Press.

Davies, D. & Bouldin, D. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2): 224–227.

Gower, J. 1985. Measures of similarity, dissimilarity, and distance. *Encyclopedia of Statistical Sciences* 5: 397–405. New York: Wiley.

Gower, J. 1986. Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification* 3:5–48.

Hartigan, J. A. (1975). Clustering Algorithms. New York: Wiley.

Kullback, S. & Leibler, R.A.. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22 (1): 79–86.

Meeker, W.Q. & Hong, Y. 2014. Reliability meets big data: Opportunities and challenges. *Quality Engineering* 26 (1): 102-116.

Moreno, P., Ho, P., N. Vasconcelos, N. 2004. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. *Proceeding of Neural Information Processing Systems*. Vancouver: Canada.

Rousseeuw, P. J. 1987. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53–65.

Von Luxburg, U. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing* 17(4): 395-416.

APPENDIX

K-means (Hartigan, 1975) is an unsupervised learning algorithm that solves the well-known clustering problem. The procedure follows a simple and easy way to partition the N vectors $(\boldsymbol{t}_1,\ldots,\boldsymbol{t}_N)$, where $\boldsymbol{t}_i=(t_1,...,t_C),\ i=1,...,N$ as defined in Eq. (11), into C clusters $S=\{S_1,\ldots,S_C\},\ \sum_{c=1}^{C}|S_c|=N$. The main idea is to define C centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each vector $\boldsymbol{t}_i,\ i=1,...,N$, and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage has done. At this point, we need to re-calculate C new centroids as barycenters of the clusters resulting from the previous step. After we have these C new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that