Computing Cohesive Force among Dataset Entities Using Differential Evolution and Hierarchical Clustering

Shahryar Rahnamayan, SMIEEE
Department of Electrical, Computer and Software Engineering
University of Ontario Institute of Technology (UOIT)
Oshawa, Canada
shahryar.rahnamayan@uoit.ca

Sedigheh Mahdavi
Department of Electrical, Computer and Software Engineering
University of Ontario Institute of Technology (UOIT)
Oshawa, Canada
sedigheh.mahdavi@uoit.ca

Kalyanmoy Deb, FIEEE
Department of Electrical and Computer Engineering
Michigan State University
East Lansing, USA
Email: kdeb@egr.msu.edu

COIN Report Number 2016020

***Abstract*—The current paper introduces a novel concept and metric, called cohesion factor (CF); which is defined pair-wise among entities in a dataset and indicates their degree of stickiness (cohesive force). First, a dataset with n entities is assumed as a whole unit or cluster, then they are clustered to $2, 3, \ldots$, and finally n classes, the more two entities are appeared in the same class during these hierarchical clustering process, the higher cohesion factor value they receive. In this direction, a new framework is proposed; which utilizes an evolutionary algorithm (i.e., DE) for optimal feature selection and a hierarchical clustering method for computing cohesion factors. The intermediate and final cohesion factors' matrices have been utilized for knowledge discovery and answering crucial data mining queries which cannot be answered by using a standalone clustering method. In order to conduct a case study, a real-world dataset is utilized; which contains 17 entities (countries) presented by corresponding 24 continuous features. The DE algorithm finds the most discriminating features in each step, which are eliminated for the next step to calculate matrix of cohesion factors. In the final step, the proposed method ranks the entities in term of their cohesion factor and features based on their elimination order. The framework is described in detail and the intermediate and final results are presented and discussed comprehensively.**

I. Introduction

Many clustering methods have been proposed to analyze various datasets. Clustering methods are analyzing tools to cluster the entities of a given dataset according to a predefined similarity measure. They are important in data mining to understand the general characteristics, relationships, and structure of the data. The various clustering methods were applied in various science and engineering applications such as biology, marketing, information retrieval, remote sensing, statistics, pattern recognition, image processing, text mining, etc. [1], [2], [3]. Generally speaking, clustering methods can be classified into two main categories; hierarchical clustering methods and partition-based clustering methods [2], [4], [5]. Hierarchical clustering methods construct a clustering tree by merging (i.e., top-down) or splitting (i.e., bottom-up) entities in the dataset. In partition-based clustering methods, clusters are constructed by reallocating entities among clusters according to the clustering criterion.

Although clustering methods are the promising means to analyze data, there is no a single clustering algorithm with the sufficient quality applicable to all real-world applications [4]. Recently, some research works have been conducted to develop automatic clustering algorithms which are able to choose a proper number of clusters for a given dataset. A large number of metaheuristic algorithms have been utilized to provide appropriate automatic clustering approaches [3], [6], [7], [8], [9], [10], [11], [12]. The metaheuristic algorithms used clustering validation criteria as an objective function to determine an optimal clustering. The efficiency of the metaheuristic algorithms is related to the basic design of their steps and operates, such as the encoding, mutation, crossover, and selection schemes [3]. Differential Evolution (DE) is one of evolutionary metaheuristic algorithms which has successfully been applied for automatic clustering algorithm [6], [13], [14], [15], [16], [3]. DE was proposed by Storn and Price in 1997 [17], [18], as one of the most effective optimization techniques.

In this paper, a new framework is proposed which uses DE and hierarchical clustering algorithms to compute a cohesion matrix of entities in dataset and ranking features based on their discriminatory abilities. The proposed framework uses DE algorithm to find the most discriminating features using an iterative procedure. After each run of DE, pair-wise cohesion factors for all entities are computed according to the hierarchical clustering tree of entities by considering only the recognized features through DE. The final cohesion matrix of all pair entities is calculated by adding up all normalized cohesion matrices obtained during all steps. The proposed framework has some major advantages in comparing to a standalone clustering method: (1) it is able to rank the features in term of their importance level to discriminate the entities of a dataset, (2) it computes pair-wise cohesion factors for the entities to analyze the dataset, (3) after eliminating a subset

of features, it is able to determine how the cohesion factor values among the entities are affected in each iteration, (4) by considering just a subset of features, it is able to indicate which clustering would be an optimal one, (5) for each entity, it finds the rankings of other entities in term of their corresponding cohesion factor values. Supporting these characteristics opens the floor to answer many valuable detailed data mining queries comprehensively. The proposed framework is applied on a dataset given in [19] which contains 17 entities (countries) described by corresponding 24 continuous features.

The organization of the rest of the paper is as follows. Section II presents a background review. Section III describes the proposed framework in details. Section IV presents the experimental results and discussion. Finally, the paper is concluded in Section V.

II. BACKGROUND REVIEW

A. Hierarchical clustering method

Hierarchical clustering algorithms construct a hierarchical tree (dendrogram) using two main approaches; agglomerative and top-down [20]. In the agglomerative mode, the algorithm starts with entities of dataset as its own cluster and then merges the most similar pair of clusters successively to build a hierarchy of clustering. Top-down mode starts with putting all entities of dataset in one cluster and then divides each cluster into smaller clusters. Sokal and Rohlf [21] introduced the Cophenetic Correlation Coefficient (CCC) which gives a measure of how faithfully the hierarchical cluster tree represents the dissimilarities among entities. The effectiveness of a hierarchical cluster tree can be evaluated by the CCC value. A higher value of CCC, closer to one, indicates a high-quality of clustering tree. CCC is defined as follow [21], [22], [23]:

$$CCC = \frac{\sum_{i<j}(x(i,j)-x)(t(i,j)-t)}{\sqrt{\sum_{i<j}(x(i,j)-x)^2 \sum_{i<j}(t(i,j)-t))^2}}, \quad (1)$$

which $x(i,j)$ is the ordinary Euclidean distance among the ith and jth entities and $t(i,j)$ is the distance between the ith and jth entities in the dendrogram.

B. Differential Evolution- A Brief Description

Storn and Price introduced a differential evolution (DE) algorithm as an effective evolutionary algorithm for optimizing the global optimization problems [17], [18]. The main operators in the classical DE (DE/1/bin) are briefly described as following.

1) Mutation operation:

$$v_i = x_{i1} + F.(x_{i2} - x_{i3}), \quad (2)$$

where r_1, r_2, r_3 are different random integer numbers within $[1, NP]$ and NP is the population size. The scaling factor F is real constant factor to control the difference vector.

2) Crossover operation:
In [17], [18], the binomial crossover is defined to generate a trial vector as follows:

$$u_{z,j} = \begin{cases} v_{z,j} & \text{rand} \leq CR \text{ or } j = j_{rand} \\ x_{z,j} & \text{otherwise} \end{cases} \quad (3)$$

where CR is the crossover rate, a constant value within the interval $[0, 1)$ and j_{rand} is a randomly number in $1, 2, \ldots, D$ and D is the problem dimension.

3) Selection scheme:
DE selects the better one between X_i and U_i according to their fitness values for the next generation (i.e., greedy selection). The selection operator chooses x'_i for the next generation in a minimization problem as follows:

$$x'_i = \begin{cases} u_i & f(u_i) < f(x_i) \\ x_i & \text{otherwise} \end{cases} \quad (4)$$

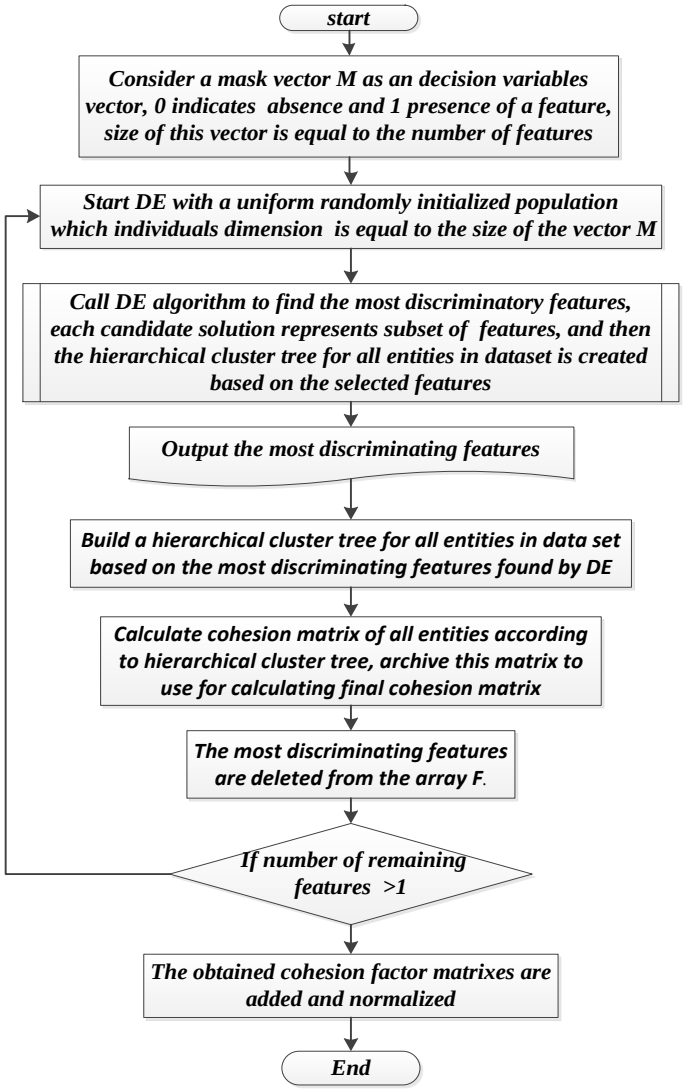


Fig. 1: The flowchart of the proposed framework.

III. THE PROPOSED FRAMEWORK

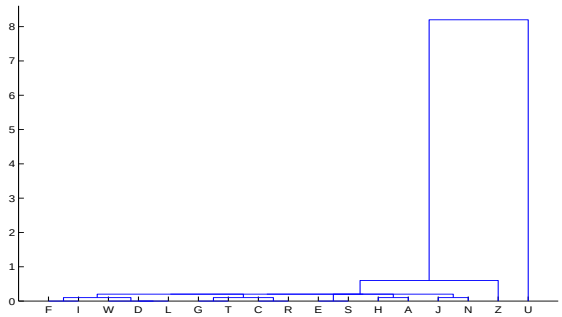
In the nature, human always seeks some methods to extract information and insights from dataset of real-world problems to recognize hidden knowledge, relationships, and properties among them. In this study, we attempt to propose a framework to obtain a cohesion matrix of entities for a dataset and a discriminatory rankings of the feature. The cohesion metric

TABLE I: Dataset used for the case study

Country	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Sweden (W)	8.6	9	9.7	9.1	9.7	9.7	9.7	7.7	4.7	10	6.7	4.3	0	0	8.7	7.5	9.2	3	9.9	7.5	5.6	4.4	5.7	1.96
Japan (J)	10	10	6.9	8.3	9.5	9.8	10	7.5	1.1	10	10	0	5.3	5.5	8.2	0	3.5	3.6	9.9	6	1.3	9.1	1.5	0.12
Denmark (D)	8.4	8	9.1	7.8	10	9.8	9.7	8	2.1	7.5	0	7.1	6	2.3	3	10	10	6.2	10	8.8	6.3	5.4	4.9	1.33
France (F)	7.3	9.2	8.3	10	9.2	8.3	9.6	7.2	0.9	7.5	0.6	9.3	1.9	3.7	1.1	2	4	3.6	6.3	2.7	8.1	6.5	6.5	2.60
Germany (G)	7.4	7.7	10	9.6	7.4	9.7	9.4	7.7	4.7	7.5	3.5	0	2.8	3.5	1.8	5	5.8	3.7	8.4	4.6	7.9	8.2	8.5	0.93
Great Britain (E)	6.9	8.9	7.6	6.1	8.8	8.6	8.8	8.9	10	5	3.5	5.7	2.6	2.7	4.7	4.5	7.9	3.1	4.8	7.1	4.2	7.2	4	1.91
Norway (N)	7.6	7	8.5	5.7	6.7	10	9.9	4.4	6.3	10	5.6	7.1	3	3.2	10	6	8.7	10	9.5	9.4	10	5.8	3.9	1.02
Holland (H)	6.2	7.4	6.4	7.5	4.9	8.8	9	8	9.6	7.5	5.1	5.7	3.7	3.7	5.5	6.5	8.3	5	7.2	6.8	9.8	8.4	4.4	0.80
Australia (A)	5.8	8.5	6.1	5.7	2.8	8.2	9.10	4.6	7.5	6	5.7	4.7	2.7	6.4	5.5	7.1	4.5	5.2	5.8	1.3	3.2	2	3	1.53
New Zealand (Z)	5.4	7.1	6.8	4.3	5.6	6.6	8.2	2.6	3.8	5	2.8	10	3	1.1	5.4	6	9.2	0.2	4.7	6.4	1.5	7	20	3.77
Canada (C)	4.1	7.3	6.4	3.5	6.2	8.3	9.3	3.6	6.2	5	7.7	2.9	2.3	4	7.1	7	8.8	5.3	6.2	6.5	1.3	5.4	17.5	5.18
Spain (S)	2.9	5.9	6.5	3	6.4	9.7	8.8	9.9	9.1	5	2.4	0	5.3	8.8	2.5	4	4.6	0	6.4	0.5	0.4	8.2	2	2.73
Switzerland (L)	5.9	6.9	5.2	3.2	4.6	9.7	9.7	2.4	0	7.5	7	1.4	2.1	6.5	3.4	10	7.7	5.6	6.2	10	5.8	7.4	19	4.34
Austria (R)	5.2	5.6	6.1	2.8	3.3	9.4	9.3	9.3	0.8	7.5	3	1.4	1.9	2.6	2.2	8	6	5.1	7.6	5.8	6.7	8.6	9	1.04
Italy (T)	2.5	1.6	5.5	1.9	5.9	8.8	9.4	10	9.4	5	5.8	0	1.6	10	5.8	3.5	0	3.3	4.8	0	4.4	10	2.3	0.42
Ireland (I)	2.2	2.7	3.1	0	5.1	9	9.6	2.5	7	2.5	1.6	8.6	2.1	8.9	0	7	5	9.8	4.9	3.7	3.5	5.8	3	1.63
United States (U)	0	0	0	1.6	0	0	0	5.6	7	0	2.3	10	10	0.1	6.9	6	6.2	9.4	0	7.9	0	0	9.9	2.81

TABLE II: Cohesion matrix among entities and Dendrogram obtained through considering just the feature 7.

(a) Dendrogram of dataset, considering just the feature 7.

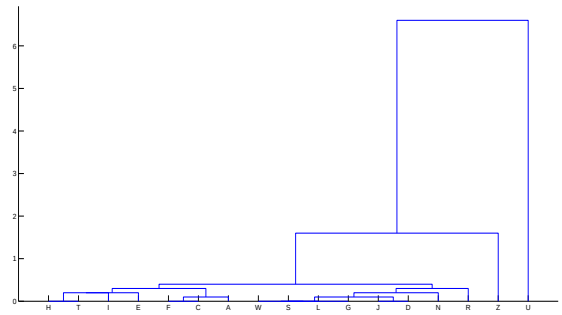


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 7.

Entity														
J	2													
D	4	2												
F	3	2	3											
G	2	2	2	2										
E	2	2	2	2	2									
N	2	3	2	2	2	2								
H	2	2	2	2	2	2	2							
A	2	2	2	2	2	2	2	3						
Z	1	1	1	1	1	1	1	1	1					
C	2	2	2	2	3	2	2	2	2	1				
S	2	2	2	2	2	4	2	2	2	1	2			
L	4	2	4	3	2	2	2	2	2	1	2	2		
R	2	2	2	2	3	2	2	2	2	1	4	2	2	
T	2	2	2	2	4	2	2	2	2	1	3	2	2	3
I	3	2	3	4	2	2	2	2	2	1	2	2	3	2	2
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE III: Cohesion matrix among entities and Dendrogram obtained through considering just the feature 6.

(a) Dendrogram of dataset, considering just the feature 6.



(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 6.

Entity														
J	5													
D	5	6												
F	2	2	2											
G	6	5	5	2										
E	2	2	2	3	2									
N	4	4	4	2	4	2								
H	2	2	2	3	2	4	2							
A	2	2	2	4	2	3	2	3						
Z	1	1	1	1	1	1	1	1	1					
C	2	2	2	5	2	3	2	3	4	1				
S	6	5	5	2	6	2	4	2	2	1	2			
L	6	5	5	2	6	2	4	2	2	1	2	6		
R	3	3	3	2	3	2	3	2	2	1	2	3	3	
T	2	2	2	3	2	4	2	5	3	1	3	2	2	2
I	2	2	2	3	2	4	2	4	3	1	3	2	2	2	2
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

ia a matrix of pair-wise stickiness measures (cohesive forces). It is a novel concept which provides more useful information to answer a wide range of data mining questions. DE and hierarchical clustering algorithms are applied to develop this framework. Fig. 1 shows an flowchart of the proposed framework; which are described in detail as following. The proposed framework is an iterative framework to calculate intermediate cohesion matrices and also final cohesion matrix. The main steps of the proposed approach are as follows: (1) running DE algorithm to find the most discriminating features; the fitness

function is given in Eq. 1, (2) creating the hierarchical clustering tree for all entities in the dataset, (3) computing a cohesion matrix for all entities by considering the recognized features found by DE, (4) eliminating the selected features found by DE from the dataset, (5) repeating steps (1) to (4) until the number of remaining feature is bigger than one, (6) calculating the cohesion matrix for the last step as before, and normalized final cohesion matrix by adding up all intermediate cohesion matrices obtained from previous steps. For a dataset D with n features $F = \{f_1, f_2, \ldots, f_n\}$, a DE algorithm is run to find the most discriminating features. DE optimizes the objective function, selecting the features which maximizes CCC value given in Eq. 1. A solution is represented as a string, length equal to n; i.e., the number of features and its each element is within $[0, 1]$. We use the classical DE but before evaluating objective function, we round each element of the candidate solution to the nearest integer, so a binary candidate solution is obtained as the mask vector. The value of the ith variable in the mask vector is '1' if the ith feature is considered as a discriminating feature, and '0' otherwise. After identifying the most discriminating features, we cluster entities in the dataset using a hierarchical clustering method. By using the hierarchical clustering tree (dendrograms), we compute the cohesion matrix of entities which is described as following. The cohesion matrix is a symmetric matrix $C = (c_{ij})$ which represents cohesion values among all possible pairs of entities in dataset. A cohesion value is a stickiness measure that how much two entities are close to each other in term of given set of features; in fact, a higher cohesion (stickiness) value means it is hard to separate two entities (because of a higher cohesive force), so they stay with each other up to a deeper stage in the created dendrogram. That is why, the cohesion factor value c_{ij} of two entities i and j is defined as the depth of the first branch of dendrogram tree at which those two entities are first joined. In the hierarchical clustering tree with depth '$0 - k$', entities are in the tree's leaves with depth k. Before beginning of the next iteration, all discriminating features are eliminated from the dataset. This process continues in the similar way for all other remaining features until there is no more feature left. Finally, the obtained cohesion matrices in all iterations are added and normalized to obtain the final cohesion matrix of all pair entities in the dataset. In each iteration of the framework, DE solves optimization problems with different dimensions. In the first iteration, the dimension of optimization problem is equal to n, the number of features, and it is reduced in the next iterations, because of elimination of the features. The pattern of feature/dimension elimination is dataset oriented, therefore, unknown in advance.

IV. EXPERIMENTAL RESULTS

A. Setup of experiments

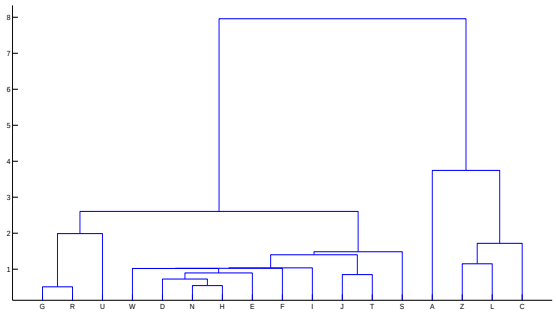
We applied the proposed method on a dataset to obtain pair-wise cohesion factors among the entities and ranking features based on their discriminating ability. The dataset used in this study are from [19] which is a well-recognized study in sociological research, it consists of 17 entities (countries) described by 24 continuous features. In [19], the competing hypotheses on the origin, mental basis, popularity, and societal efficacy of mass religion versus secularism were investigated and analyzed among 17 countries using the Pearson correlations. We select features of dataset which they do not have a missing value for any entity. These features are absolutely believe in god (1), bible literalists (2), prayer (3), agnostics and atheists (4), acceptance of evolution (5), homicides (6), incarceration (7), suicide 15-24 (8), suicide all age (9), Under 5 mortality (10), lifespan (11), fertility (12), marriages (13), divorces (14), alcohol consumption (15), life satisfaction (16), corruption (17), per capita income (18), income inequality (19), employment levels (20), average hours worked (21), resource exploitation base (22), foreign born (23), and cultural fractionalization (24). The entities of the dataset are the following countries: Sweden (W), Japan (J), Denmark (D), France (F), Germany (G), Great Britain (E), Norway (N), Holland (H), Australia (A), New Zealand (Z), Canada (C), Spain (S), Switzerland (L), Austria (R), Italy (T), Ireland (I), and United States (U). Table I presents the utilized dataset. For the DE algorithm, the population size was set to 100, the number of generation was set to 20,000 as the terminating condition, crossover rate (CR) and mutation factor are set to 0.9 and 0.5, respectively.

B. Numerical Results

The time complexity of the proposed method depends on the given problem, how fast the features are eliminated during the iterations. For this dataset, DE is run for 14 times to obtain cohesion matrices among entities. In the step 1, DE is run for the problem with 24 variables (all variables) and it finds the feature 7 (Incarceration) as the most discriminating feature with the CCC value of 0.9938. Table II shows a dendrogram plot of the hierarchical cluster tree which is obtained by only considering of the feature 7 and corresponding cohesion matrix. Step 2 runs DE with 23 variables (now the feature 7 is eliminated from the dataset) and this time it finds the feature 6 (Homicides) as the most discriminating feature, with the CCC value of 0.9829. Table III shows its dendrogram plot which is obtained through considering the feature 6 and corresponding cohesion matrix. For steps 3 to 14, the discriminating features are found by DE (as listed in Table XVII) with the given CCC values, and corresponding dendrograms and cohesion matrix are given in Table IV to Table XV, respectively. Finally, all obtained cohesion matrices are added and normalized to obtain the final cohesion matrix of all entities in dataset. Table XVI shows the final cohesion matrix of all entities. In the proposed algorithm, we can rank features based on their discrimination ability. The rank of each feature is its order according to feature selection steps in the process of algorithm, if f_i is selected sooner than f_j, it means feature f_i has a higher discrimination power (let say it separates entities more stronger). Table XVII shows the ranks of features and the order of their selection. The proposed framework can address the following kind of questions: 1) Considering a specific country, what is the ranking of other countries which are closer to this country in term of given set of features? 2) Considering a subset of the countries and features, what is the ranking of the features in term of their power of discrimination? 3) What is the ranking of pair-wise closeness of countries based on the given set of features? 4) How eliminating of the feature(s) can affect pair-wise closeness of countries? 5) How eliminating of the subset of countries can affect pair-wise closeness of the remaining countries? None of these questions cannot be answered just by applying a standalone clustering method. Extracting from Table XVI, Table XVIII is obtained, which presents pair-wise ranking of the countries based on their

TABLE IV: Dendrogram and cohesion matrix among entities obtained through the features 23 and 24.

(a) Dendrogram of dataset, considering just the features 23 and 24.

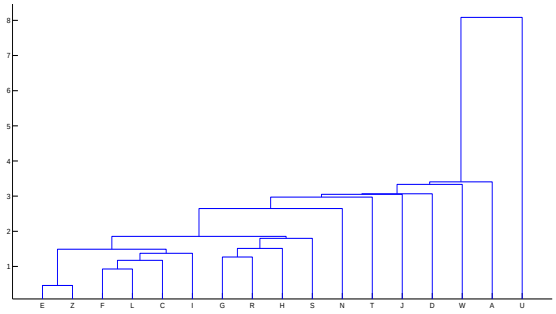


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the features 23 and 24.

Entity															
J	3														
D	4	3													
F	4	3	4												
G	1	1	1	1											
E	4	3	5	4	1										
N	4	3	6	4	1	5									
H	4	3	6	4	1	5	7								
A	0	0	0	0	0	0	0	0							
Z	0	0	0	0	0	0	0	0	1						
C	0	0	0	0	0	0	0	0	1	2					
S	2	2	2	2	1	2	2	2	0	0	0				
L	0	0	0	0	0	0	0	0	1	3	2	0			
R	1	1	1	1	3	1	1	1	0	0	0	1	0		
T	3	4	3	3	1	3	3	3	0	0	0	1	0	1	
I	4	3	4	4	1	4	4	4	0	0	0	1	0	1	3
U	1	1	1	1	2	1	1	1	0	0	0	1	2	1	1
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE V: Dendrogram and cohesion matrix among entities obtained through the features 13, 22, and 19.

(a) Dendrogram of dataset, considering just the features 13, 22, and 19.

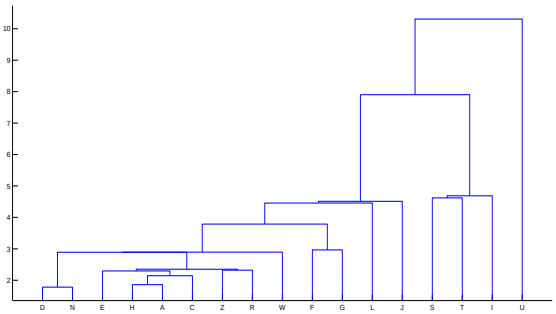


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the features 13, 22, and 19.

Entity															
J	2														
D	2	3													
F	2	3	3												
G	2	3	3	6											
E	2	3	3	7	6										
N	2	3	3	5	5	5									
H	2	3	3	6	8	6	5								
A	1	1	1	1	1	1	1	1							
Z	2	3	3	7	6	8	5	6	1						
C	2	3	3	9	6	7	5	6	1	7					
S	2	3	3	6	7	6	5	7	1	6	6				
L	2	3	3	10	6	7	5	6	1	7	9	6			
R	2	3	3	6	9	6	5	8	1	6	6	7	6		
T	2	3	3	4	4	4	4	4	1	4	4	4	4	4	
I	2	3	3	8	6	7	5	6	1	7	8	6	8	6	4
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE VI: Dendrogram and cohesion matrix among entities obtained through the features 1, 2, 3, 14, and 21.

(a) Dendrogram of dataset, considering the features 1, 2, 3, 14, and 21.



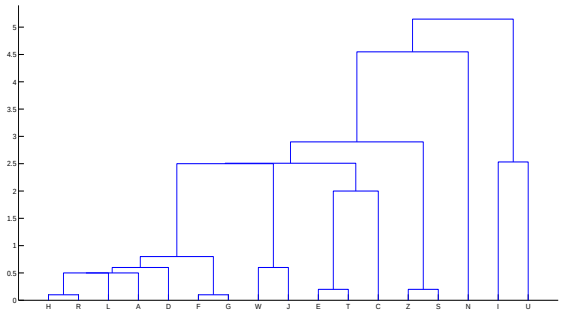
(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the features 1, 2, 3, 14, and 21.

Entity															
J	2														
D	5	2													
F	4	2	4												
G	4	2	4	5											
E	5	2	5	4	4										
N	5	2	5	4	4	5									
H	5	2	5	4	4	7	5								
A	5	2	5	4	4	7	5	9							
Z	5	2	5	4	4	6	5	6	6						
C	5	2	5	4	4	7	5	8	8	6					
S	1	1	1	1	1	1	1	1	1	1	6				
L	3	2	3	3	3	3	3	3	3	3	3	1			
R	5	2	5	4	4	6	5	6	6	7	6	1	3		
T	1	1	1	1	1	1	1	1	1	1	1	3	1	1	
I	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

cohesion values. They are in 37 groups (some pairs has the same cohesion value). The Austria-Germany has the highest value (i.e., CF=0.89); United States-Japan and United States-Denmark have the lowest value (i.e., CF=0). The cohesion

TABLE VII: Dendrogram and cohesion matrix among entities obtained through obtained through the features 10 and 18.

(a) Dendrogram of dataset, considering just the features 10 and 18.

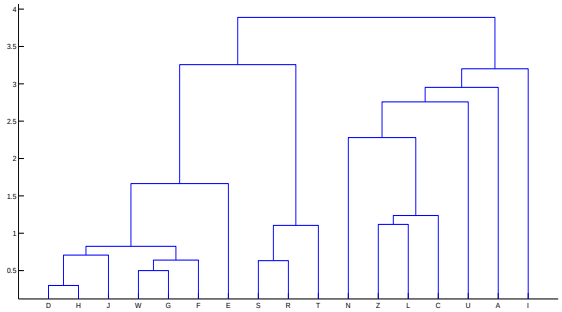


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the features 10 and 18.

Entity															
J	4														
D	3	3													
F	3	3	5												
G	3	3	5	5											
E	3	3	3	3	3										
N	1	1	1	1	1	1									
H	3	3	5	4	4	3	1								
A	3	3	5	4	4	3	1	6							
Z	2	2	2	2	2	2	1	2	2						
C	3	3	3	3	3	4	1	3	3	2					
S	2	2	2	2	2	2	1	2	2	3	2				
L	2	2	5	4	4	5	1	6	6	2	3	2			
R	3	3	5	4	4	5	1	7	6	2	3	2	6		
T	3	3	3	3	3	5	1	3	3	2	4	2	3	3	
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE VIII: Dendrogram and cohesion matrix among entities obtained through the features 4 and 8.

(a) Dendrogram of dataset, considering just the features 4 and 8.

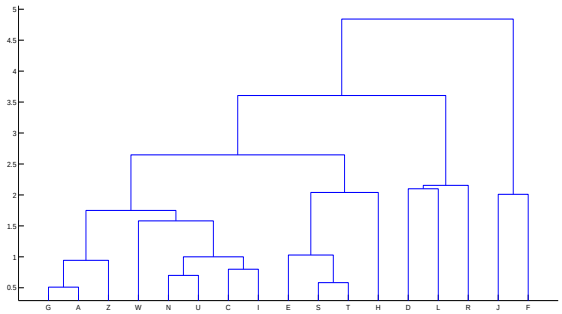


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the features 4 and 8.

Entity															
J	3														
D	3	4													
F	4	3	3												
G	5	3	3	4											
E	2	2	2	2	2										
N	0	0	0	0	0	0									
H	3	4	5	3	3	2	0								
A	0	0	0	0	0	0	2	0							
Z	0	0	0	0	0	0	4	0	2						
C	0	0	0	0	0	0	4	0	2	5					
S	1	1	1	1	1	1	0	1	0	0	0				
L	0	0	0	0	0	0	4	0	2	6	5	0			
R	1	1	1	1	1	1	0	1	0	0	0	3	0		
T	1	1	1	1	1	1	0	1	0	0	0	2	0	2	
I	0	0	0	0	0	0	1	0	1	1	1	0	1	0	0
U	0	0	0	0	0	0	3	0	2	3	3	0	3	0	0	1
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE IX: Dendrogram and cohesion matrix among entities obtained through the features 9 and 16.

(a) Dendrogram of dataset, considering just the features 9 and 16.



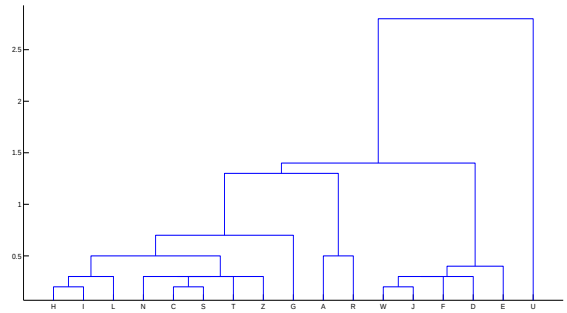
(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the features 9 and 16.

Entity															
J	0														
D	1	0													
F	0	1	1												
G	3	0	1	0											
E	2	0	1	0	3										
N	4	0	1	0	3	2									
H	2	0	1	0	2	3	2								
A	3	0	1	0	5	3	3	3							
Z	3	0	1	0	4	3	3	3	4						
C	4	0	1	0	3	2	5	2	3	3					
S	2	0	1	0	2	4	2	3	2	2	2				
L	1	0	3	0	1	1	1	1	1	1	1	1			
R	1	0	2	0	1	1	1	1	1	1	1	1	2		
T	2	0	1	0	2	4	2	3	2	2	2	5	2	2	
I	4	0	1	0	3	2	5	2	3	3	6	2	2	1	1
U	4	0	1	0	3	2	6	2	3	3	5	2	2	1	1	5
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

value plot for all possible pairs (i.e., 136) are given in Fig. 2, ranging form 0 to 0.89. Table XVII presents the ranking of the features in term of their discriminating abilities, as seen, the feature 7 (i.e., incarceration) has the highest ranking and the

TABLE X: Dendrogram and cohesion matrix among entities obtained through the feature 5.

(a) Dendrogram of dataset, considering just the feature 5.

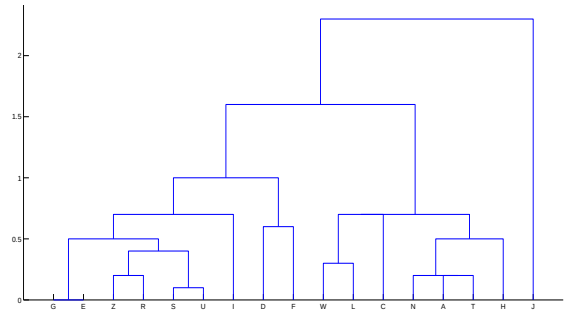


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 5.

Entity															
J	4														
D	3	3													
F	3	3	3												
G	1	1	1	1											
E	2	2	2	2	1										
N	1	1	1	1	3	1									
H	1	1	1	1	3	1	4								
A	1	1	1	1	2	1	2	2							
Z	1	1	1	1	3	1	5	4	2						
C	1	1	1	1	3	1	4	4	2	5					
S	1	1	1	1	3	1	4	4	2	5	6				
L	1	1	1	1	3	1	4	5	2	4	4	4			
R	1	1	1	1	2	1	2	2	3	2	2	2	2		
T	1	1	1	1	3	1	5	4	2	5	5	5	4	2	
I	1	1	1	1	3	1	4	6	2	4	4	4	5	2	4
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE XI: Dendrogram and cohesion matrix among entities obtained through the feature 11.

(a) Dendrogram of dataset, considering just the feature 11.

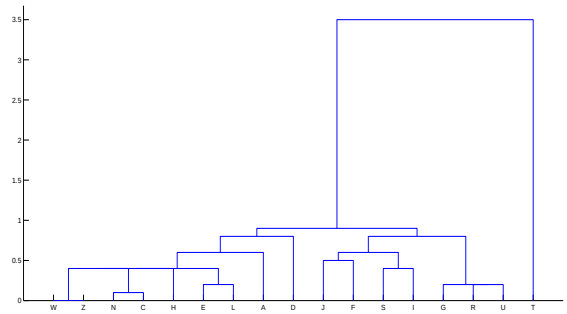


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 11.

Entity															
J	0														
D	1	0													
F	1	0	3												
G	1	0	2	2											
E	1	0	2	2	5										
N	2	0	1	1	1	1									
H	2	0	1	1	1	1	3								
A	2	0	1	1	1	1	4	3							
Z	1	0	2	2	4	4	1	1	1						
C	2	0	1	1	1	1	2	2	2	1					
S	1	0	2	2	4	4	1	1	1	5	1				
L	3	0	1	1	1	1	2	2	2	1	2	1			
R	1	0	2	2	4	4	1	1	1	6	1	5	1		
T	2	0	1	1	1	1	4	3	4	1	2	1	1	1	
I	1	0	2	2	3	3	1	1	1	3	1	3	2	3	1
U	1	0	2	2	4	4	1	1	1	5	1	6	1	5	1	3
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE XII: Dendrogram and cohesion matrix among entities obtained through the feature 17.

(a) Dendrogram of dataset, considering just the feature 17.



(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 17.

Entity															
J	1														
D	2	1													
F	1	4	1												
G	1	2	1	2											
E	4	1	2	1	1										
N	5	1	2	1	1	4									
H	4	1	2	1	1	4	4								
A	3	1	2	1	1	3	3	3							
Z	5	1	2	1	1	4	4	4	3						
C	5	1	2	1	1	4	5	4	3	4					
S	1	3	1	3	2	1	1	1	1	1	1				
L	4	1	2	1	1	5	4	4	3	4	4	1			
R	1	2	1	2	4	1	1	1	1	1	1	2	1		
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
I	1	3	1	3	2	1	1	1	1	1	1	4	1	2	0
U	1	2	1	2	4	1	1	1	1	1	1	2	1	4	0	2
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

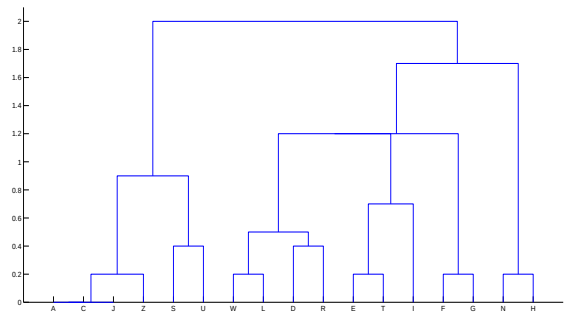
feature 12 (i.e., fertility) has the lowest ranking.

V. Conclusion

In this paper, we proposed a new framework to calculate a cohesion matrix of entities in a dataset and also to obtain

TABLE XIII: Dendrogram and cohesion matrix among entities obtained through the feature 20.

(a) Dendrogram of dataset, considering just the feature 20.

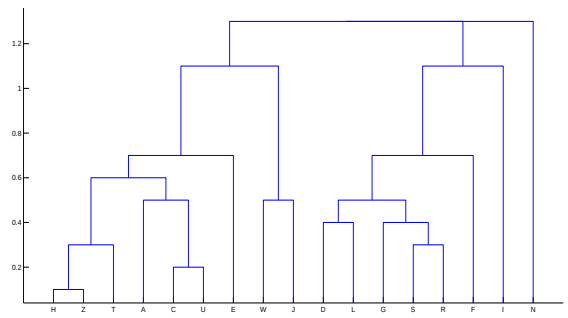


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 20.

Entity															
J	0														
D	3	0													
F	2	0	2												
G	2	0	2	3											
E	2	0	2	2	2										
N	1	0	1	1	1	1									
H	1	0	1	1	1	1	2								
A	0	3	0	0	0	0	0	0							
Z	0	2	0	0	0	0	0	0	2						
C	0	3	0	0	0	0	0	0	3	2					
S	0	1	0	0	0	0	0	0	1	1	1				
L	4	0	3	2	2	2	1	1	0	0	0	0			
R	3	0	4	2	2	2	1	1	0	0	0	0	3		
T	2	0	2	2	2	4	1	1	0	0	0	0	2	2	
I	2	0	2	2	2	3	1	1	0	0	0	0	2	2	3
U	0	1	0	0	0	0	0	0	1	1	1	2	0	0	0	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE XIV: Dendrogram and cohesion matrix among entities obtained through the feature 15.

(a) Dendrogram of dataset, considering just the feature 15.

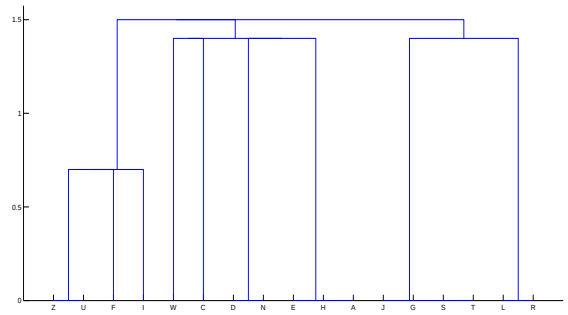


(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 15.

Entity															
J	2														
D	0	0													
F	0	0	2												
G	0	0	3	2											
E	1	1	0	0	0										
N	0	0	0	0	0	0									
H	1	1	0	0	0	2	0								
A	1	1	0	0	0	2	0	3							
Z	1	1	0	0	0	2	0	5	3						
C	1	1	0	0	0	2	0	3	4	3					
S	0	0	3	2	4	0	0	0	0	0	0				
L	0	0	4	2	3	0	0	0	0	0	0	3			
R	0	0	3	2	4	0	0	0	0	0	0	5	3		
T	1	1	0	0	0	2	0	4	3	4	3	0	0	0	
I	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0
U	1	1	0	0	0	2	0	3	4	3	5	0	0	0	3	0
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE XV: Dendrogram and cohesion matrix among entities obtained through the feature 12.

(a) Dendrogram of dataset, considering just the feature 12.



(b) Cohesion matrix among entities for the hierarchical cluster tree, obtained through the feature 12.

Entity															
J	0														
D	1	0													
F	0	0	0												
G	0	2	0	0											
E	1	0	0	0	0										
N	1	0	2	0	0	1									
H	1	0	1	0	0	2	1								
A	1	0	0	0	0	2	1	2							
Z	0	0	0	1	0	0	0	0	0						
C	1	0	1	0	0	1	1	1	1	0					
S	0	2	0	0	2	0	0	0	0	0	0				
L	0	1	0	0	1	0	0	0	1	0	0	1			
R	0	1	0	0	1	0	0	0	1	0	0	1	2		
T	0	2	0	0	2	0	0	0	0	0	0	2	0	2	
I	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
U	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	1
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

ranking of the features based on their discriminating abilities. . The proposed framework is an iterative method which runs DE algorithm to find the most discriminating features step-by-step. The cophenetic correlation coefficient (CCC) of a hierarchical

TABLE XVIII: Ranking of pair-wise cohesion values for all 136 possibilities, categorized in 37 groups; some of the pairs have the same cohesion value.

Rank:CF	1:0.84	2:0.82	3:0.76	4:0.73	5:0.71	6:0.69
Country	Austria-Germany	Holland-Great Britain	Canada-New Zealand	Canada-Australia	Holland-Norway Australia-Holland Canada-Holland Denmark-Sweden Spain-Germany Switzerland-Canada	France-Denmark Canada-Norway
Rank:CF	7:0.67	8:0.64	9:0.62	10:0.60	11:0.58	12:0.56
Country	Germany-France Holland-Denmark Austria-Spain	Canada-Great Britain Switzerland-Denmark Austria-Switzerland Italy-Holland	Germany-Denmark Great Britain-Sweden Holland-Sweden New Zealand-Holland Switzerland-Germany Switzerland-New Zealand Austria-Denmark Austria-Holland	Great Britain-France Great Britain-Germany Norway-Sweden Holland-Germany New Zealand-Great Britain Switzerland-Holland Austria-Great Britain Italy-Great Britain	Germany-Sweden Great Britain-Denmark Switzerland-Sweden Switzerland-Norway	Norway-Denmark Norway-Great Britain Holland-France New Zealand-Norway Ireland-France
Rank:CF	13:0.53	14:0.51	15:0.49	16:0.47	17:0.46	18:0.44
Country	France-Sweden New Zealand-Australia Switzerland-France Switzerland-Great Britain Austria-France Italy-Spain	Japan-Sweden Australia-Great Britain Canada-Sweden Spain-Great Britain Switzerland-Spain Ireland-Great Britain Ireland-Holland Ireland-Switzerland	Denmark-Japan Austria-New Zealand Italy-Canada Ireland-Norway Ireland-Canada Ireland-Spain	France-Japan Norway-Germany New Zealand-Germany Canada-France Canada-Germany Spain-Holland Spain-New Zealand Switzerland-Australia Austria-Canada Italy-Germany Ireland-Germany	Australia-Norway Italy-Norway	Italy-Austria
Rank:CF	19:0.42	20:0.40	21:0.38	22:0.36	23:0.33	24:0.31
Country	Germany-Japan Australia-Sweden Spain-Denmark Spain-France Spain-Canada Austria-Sweden Austria-Australia	Spain-Japan Spain-Norway Austria-Norway Ireland-New Zealand Ireland-Austria	Norway-France Holland-Japan Australia-Germany New Zealand-Sweden Italy-Sweden Ireland-Italy	Great Britain-Japan Canada-Denmark Spain-Sweden Italy-France Italy-Australia Italy-New Zealand Italy-Switzerland Ireland-Sweden Ireland-Denmark	Australia-Denmark New Zealand-France Italy-Japan Italy-Denmark	Austria-Japan
Rank:CF	25:0.29	26:0.24	27:0.22	28:0.20	29:0.18	30:0.17
Country	Norway-Japan Australia-France New Zealand-Denmark Canada-Japan Switzerland-Japan United States-New Zealand	Australia-Japan United States-Canada	Spain-Australia Ireland-Japan Ireland-Australia	New Zealand-Japan United States-Ireland	United States-Spain	United States-Germany
Rank:CF	31:0.16	32:0.15	33:0.11	34:0.06	35:0.04	36:0.02
Country	United States-Norway United States-Austria	United States-Australia	United States-Great Britain	United States-Sweden United States-Holland	United States-Switzerland	United States-France United States-Italy
Rank:CF	37:0					
Country	United States-Japan United States-Denmark					

TABLE XVI: The final cohesion factors among entities.

Entity																
J	0.51															
D	0.71	0.49														
F	0.53	0.47	0.69													
G	0.58	0.42	0.62	0.67												
E	0.62	0.36	0.58	0.60	0.60											
N	0.60	0.29	0.56	0.38	0.47	0.56										
H	0.62	0.38	0.67	0.56	0.60	0.84	0.73									
A	0.42	0.24	0.33	0.29	0.38	0.51	0.46	0.73								
Z	0.38	0.20	0.29	0.33	0.47	0.60	0.56	0.62	0.53							
C	0.51	0.29	0.36	0.47	0.47	0.64	0.69	0.73	0.76	0.82						
S	0.36	0.40	0.42	0.42	0.71	0.51	0.40	0.47	0.22	0.47	0.42					
L	0.58	0.29	0.64	0.53	0.62	0.53	0.58	0.60	0.47	0.62	0.71	0.51				
R	0.42	0.31	0.62	0.53	0.89	0.60	0.40	0.62	0.42	0.49	0.47	0.67	0.64			
T	0.38	0.33	0.33	0.36	0.47	0.60	0.44	0.64	0.36	0.36	0.49	0.53	0.36	0.44		
I	0.36	0.22	0.36	0.56	0.47	0.51	0.49	0.51	0.22	0.40	0.49	0.49	0.51	0.40	0.38	
U	0.06	0	0	0.02	0.17	0.11	0.16	0.06	0.15	0.29	0.24	0.18	0.04	0.16	0.02	0.20
Entity	W	J	D	F	G	E	N	H	A	Z	C	S	L	R	T	I

TABLE XVII: The ranks of features and the order of their selection.

Steps of algorithm	Selected feature(s) by DE	Rank	CCC
1	7	1	0.99
2	6	2	0.98
3	23,24	3	0.98
4	13,22,19	4	0.93
5	1,2,3,14,21	5	0.91
6	10,18	6	0.89
7	4,8	7	0.82
8	9,16	8	0.82
9	5	9	0.81
10	11	10	0.80
11	17	11	0.77
12	20	12	0.75
13	15	13	0.70
14	12	14	0.55

clustering tree is used as the objective function for the DE. In each iteration, a cohesion matrix for all entities, by considering only the recognized features, is computed. The final step of the proposed framework computes the final cohesion factors

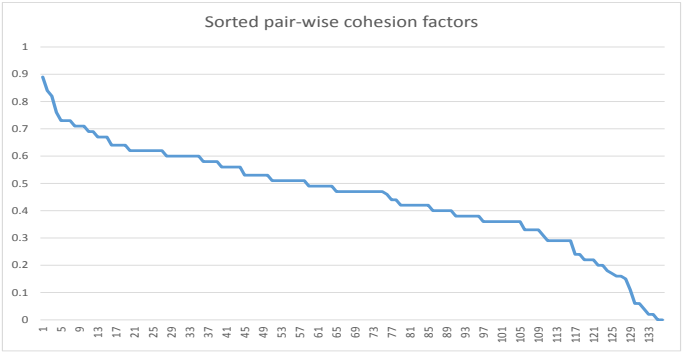


Fig. 2: The sorted arrangements of cohesion factors.

of all pair entities in dataset by adding and normalizing the obtained cohesion matrices obtained in all iterations. This framework conducts a systematic and simple approach by employing DE and hierarchical clustering algorithm to advance the investigation and analysis of a dataset. The outputs of the framework enable us to answer very detail questions regarding the entities and features, which are not possible to handle by applying a standalone clustering approach. In the paper, the proposed framework is applied to a real-world dataset with 17 entities (countries) described by 25 continuous features. In future, we are planning to develop strategies based on the proposed framework to cover a wider range of the data mining queries. In addition, we are interested in applying the proposed framework to more large-scale dataset.

REFERENCES

[1] S. Theodoridis and K. Koutroumbas, "Pattern recognition," 2003.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[3] A. José-García and W. Gómez-Flores, "Automatic clustering using nature-inspired metaheuristics: A survey," *Applied Soft Computing*, vol. 41, pp. 192–213, 2016.

[4] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[5] R. Xu, D. Wunsch *et al.*, "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp. 645–678, 2005.

[6] S. Das, A. Abraham, and A. Konar, "Metaheuristic pattern clustering–an overview," in *Metaheuristic Clustering*. Springer, 2009, pp. 1–62.

[7] H. He and Y. Tan, "A two-stage genetic algorithm for automatic clustering," *Neurocomputing*, vol. 81, pp. 49–59, 2012.

[8] C.-W. Bong and M. Rajeswari, "Multi-objective nature-inspired clustering and classification techniques for image segmentation," *Applied Soft Computing*, vol. 11, no. 4, pp. 3271–3282, 2011.

[9] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 1, pp. 56–76, 2007.

[10] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of clusters and application to image classification," *Pattern Recognition*, vol. 35, no. 6, pp. 1197–1208, 2002.

[11] E. R. Hruschka, R. J. Campello, A. Freitas, A. C. De Carvalho *et al.*, "A survey of evolutionary algorithms for clustering," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 2, pp. 133–155, 2009.

[12] S. J. Nanda and G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," *Swarm and Evolutionary Computation*, vol. 16, pp. 1–18, 2014.

[13] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 1, pp. 218–237, 2008.

[14] S. Das and S. Sil, "Kernel-induced fuzzy clustering of image pixels with an improved differential evolution algorithm," *Information Sciences*, vol. 180, no. 8, pp. 1237–1256, 2010.

[15] U. Maulik and S. Bandyopadhyay, "Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 41, no. 5, pp. 1075–1081, 2003.

[16] U. Maulik and I. Saha, "Automatic fuzzy clustering using modified differential evolution for image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 9, pp. 3503–3510, 2010.

[17] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[18] ——, *Differential evolution-a simple and efficient adaptive scheme for global optimization over continuous spaces*. ICSI Berkeley, 1995, vol. 3.

[19] G. Paul, "The chronic dependence of popular religiosity upon dysfunctional psychosociological conditions," *Evolutionary Psychology*, vol. 7, no. 3, pp. 398–441, 2009.

[20] C. Shalizi, "Distances between clustering, hierarchical clustering," *Lectures notes*, 2009.

[21] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, pp. 33–40, 1962.

[22] J. S. Farris, "On the cophenetic correlation coefficient," *Systematic Biology*, vol. 18, no. 3, pp. 279–285, 1969.

[23] S. Saraçli, N. Doğan, and İ. Doğan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation," *Journal of Inequalities and Applications*, vol. 2013, no. 1, pp. 1–8, 2013.