STUDIA UNIV. "BABEŞ-BOLYAI", INFORMATICA, Volume XLIII, Number 1, 1998

A GENETIC ALGORITHM FOR SYMBOLIC DATA CLUSTERING

DAN DUMITRESCU BEATRICE LAZZERINI
ANTON LEHENE

1. Introduction

The aim of this paper is to propose an **evolutionary** approach for the symbolic data clustering.

The objective of cluster analysis is to group a set of objects into clusters such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity.

2. Symbolic Objects

For the definition for symbolic objects, and for distance measures, we follow the definitions given by Diday [2] and Gowda and Diday [6], [7].

Let A be a symbolic object having **s** characteristics. The object A can be represented as a vector with **s** components such as:

$$A = (A_1, A_2, ..., A_s),$$

where s is the number of characteristics of object, and A_j is the value for the j-th characteristic, $j = 1, ..., s$.

Let us recall the example given in [8], example that is presented in Table I. In this table are described twelve objects (computers), with five characteristics: DISPLAY, RAM, ROM, MP, Keys. The names of the symbolic objects are in the "MICROCOMPUTER" column. In the case of Table I, $s = 5$ (number of characteristics). As an example for the object number 3 in Table I, which has the name "Exidi sorcerer" we have the following values of the characteristics:

Characteristic DISPLAY having the value (A_1) B&W TV;

Characteristic RAM having the value (A_2) 48K;

Characteristic ROM having the value (A_3) 4K;

Characteristic Mp having the value (A_4) Z80;

Characteristic Keys having the value (A_5) 57-63.

Then, conform to the definition given before, the object "Exidi sorcerer", can be written like this:

Exidi sorcerer = (B&W TV, 48K, 4K, Z80, (57-63)).

1991 *Mathematics Subject Classification.* 68H30, 68T99.
1991 *CR Categories and Subject Descriptors.* I.5.3 [**Pattern Recognition**]: Clustering - *algorithms, similarity measures.*

The characteristic values may be measured on different scales, resulting in the following types:

1. quantitative characteristics, which can be classified into continuous, discrete and interval values;
2. qualitative characteristics, which can be classified into nominal(unordered), ordinal(ordered), and combinational.

For the objects given in the Table I, characteristics DISPLAY and MP are qualitative and characteristics: RAM, ROM, Keys are quantitative.

TABLE I: Microcomputer Data

MICROCOMPUTER	*DISPLAY*	*RAM*	*ROM*	*MP*	*Keys*
0 Apple II	COLOR TV	48K	10K	6502	52
1 Atari 800	COLOR TV	48K	10K	6502	57-63
2 Commodore VIC 20	COLOR TV	32K	11-16K	6502A	64-73
3 Exidi sorcerer	B&W TV	48K	4K	Z80	57-63
4 Zenith H8	BUILT-IN	64K	1K	8080A	64-73
5 Zenith H89	BUILT-IN	64K	8K	Z80	64-73
6 HP-85	BUILT-IN	32K	80K	HP	92
7 Horizon	TERMINAL	64K	8K	Z80	57-63
8 Sc. Challenger	B&W TV	32K	10K	6502	53-56
9 Ohio Sc. II Series	B&W TV	48K	10K	6502C	53-56
10 TRS-80 I	B&W TV	48K	12K	Z80	53-56
11 TRS-80 III	BUILT-IN	48K	14K	Z80	64-73

3. Dissimilarity measures

The distance measure that we will use is the distance measure introduced by Gowda and Diday [6].

Suppose we have two symbolic objects A and B:

$$A = (A_1, A_2, ..., A_s), \quad B = (B_1, B_2, ..., B_s)$$

For the quantitative type of the k-th characteristic, $k = 1, ..., s$, we define **the length of maximum interval (k-th characteristic)** as the difference between the highest and the lowest values of the k-th characteristic over all objects.

We define **the span length of A_k and B_k** as the length of minimum interval containing both A_k and B_k.

For the qualitative type of the k-th characteristic, k = 1,...,**s**, we define the **length of a characteristic value** as the number of its elements.

And we define **the span length of A_k and B_k** is defined as the number of the elements in their union.

In the definition of the next notions we will use symbolic objects A and B, to illustrate these notions.

For the k-th characteristic, k=1,...,**s** we have the following dissimilarity measures between A_k and B_k :

a) Quantitative type of the k-th characteristic:

1) $D_p(A_k, B_k)$, the dissimilarity measure due to position p. D_p is defined as:

$D_p(A_k, B_k)$ = |lower limit of A_k – lower limit of B_k | / (length of maximum interval (k-th characteristic));

2). $D_S(A_k, B_k)$, the dissimilarity measure due to span s. D_S is defined as:

$D_S(A_k, B_k)$ = |length of A_k – length of B_k | / (span length of A_k and B_k);

3). $D_C(A_k, B_k)$, the dissimilarity measure due to content c. D_C is defined as:

$D_C(A_k, B_k)$ = (|length of A_k + length of B_k – 2·length of intersection between A_k and B_k |) / (span length of A_k and B_k).

b) Qualitative type of the k-th characteristic:

1). $D_S(A_k, B_k)$, the dissimilarity measure due to span s, defined as:

$D_S(A_k, B_k)$ = |length of A_k – length of B_k | / (span length of A_k and B_k);

2). $D_C(A_k, B_k)$, the dissimilarity measure due to content c, defined as:

$D_C(A_k, B_k)$ = (|length of A_k + length of B_k – 2·length of intersection between A_k and B_k |) / (span length of A_k and B_k).

We observe that the dissimilarity measure D_p due to position p is defined for the quantitative characteristics only.

The total dissimilarity measure between A_k and B_k when the k-th characteristic is quantitative is:

(1) $$D(A_k, B_k) = D_p(A_k, B_k) + D_s(A_k, B_k) + D_c(A_k, B_k).$$

The dissimilarity measure when the k-th characteristic is qualitative is defined as:

(2) $$D(A_k, B_k) = D_s(A_k, B_k) + D_c(A_k, B_k).$$

Remarks:

1. If the values for the k-th characteristic A_k and B_k are the same ($A_k = B_k$) then $D(A_k, B_k) = 0$, regardless of the type of the k-th characteristic.
2. The dissimilarity measures D_p , D_S , D_C are normalized between zero and one.

Now we can define the dissimilarity measure between two symbolic objects A and B as:

$$D(A, B) = \sum_{k=1}^{s} D(A_k, B_k),$$

where s represents the number of characteristics.

Example: Assume we would like to calculate the dissimilarity between the objects with number two (Commodore VIC 20) and eight (SC. Challenger).

D(Commodore VIC 20, Sc. Challenger) =
D(COLOR TV, B&W TV) + D(32K, 32K) + D(11-16K, 10K)
+ D(6502A, 6502) + D(64-73, 53-56).

Let us calculate the dissimilarity measures for the qualitative characteristics. For the DISPLAY characteristic we have the following dissimilarity measure:

D(COLOR TV, B&W TV) =
D_S (COLOR TV, B&W TV) + D_C (COLOR TV, B&W TV) = 0 + 1 = 1,

which is composed from:

D_S (COLOR TV, B&W TV) = |1–1| / 2 = 0,

D_C (COLOR TV, B&W TV) = |1+1−2·0| / 2 = 1.

Therefore the dissimilarity measure for the DISPLAY characteristic is:

D(COLOR TV, B&W TV) = 0 + 1 = 1.

For the MP characteristic we have:

$$D(6502A, 6502) = D_S(6502A, 6502) + D_C(6502A, 6502)$$
$$= |1-1| / 2 + |1+1-2\cdot 0| / 2 = 0 + 1 = 1.$$

Let us now calculate the dissimilarity measures for the quantitative characteristics:

For the RAM characteristic we have the following dissimilarity measure:

$$D(32K, 32K) = D_P(32K, 32K) + D_S(32K, 32K) + D_C(32K, 32K) = 0,$$

which is composed from:

$$D_P(32K, 32K) = 0,\ D_S(32K, 32K) = 0,\ D_C(32K, 32K) = 0.$$

For the ROM characteristic we have:

$$D(11\text{-}16K, 10K) = D_P(11-16K, 10K) + D_S(11-16K, 10K) + D_C(11-16K, 10K)$$
$$= 0.0126 + 0.833 + 0.833 = 1.6786,$$

which is composed from:

$$D_P(11\text{-}16K, 10K) = |11 - 10| / (80 - 1) = 1 / 79 = 0.0126,$$

where 80 – 1 is the maximum length of the interval, for the "ROM" characteristic,

$$D_S(11-16K, 10K) = |5 - 0| / (16 - 10) = 5 / 6 = 0.8333,$$

$$D_C(11-16K, 10K) = |5 + 0 - 2\cdot 0| / (16 - 10) = 5 / 6 = 0.8333,$$

where 16 – 10 is the length of minimum interval containing both 10K and 11 – 16K.

For the Keys characteristic we have:

$$D(64\text{-}73, 53\text{-}56) = D_P(64\text{-}73, 53\text{-}56) + D_S(64\text{-}73, 53\text{-}56) + D_C(64\text{-}73, 53\text{-}56)$$
$$= 0.275 + 0.3 + 0.6 = 1.175.$$

Therefore we have:

D(Commodore VIC 20, Sc. Challenger) = 1 + 0 + 1.6786 + 1 + 1.175 4.8536.

To formulate a clustering problem for symbolic data we have to define the **prototype of a cluster** and a **distance from a symbolic object to a prototype**.

Consider a set of symbolic objects: $X = \{X^1, X^2, ..., X^p\}$. The problem is to detect **n** optimal clusters in X.

For the definition of a **prototype** and a **distance measure between a symbolic object and a prototype** we follow the definition given by Yasser and Ismail in [8].

Let us consider that the cluster structure of our data set X is described by a fuzzy partition $P = \{A_1, A_2, ..., A_n\}$ of X. In this case we denote with $A_i(X^j)$ the membership degree of the object X^j to the cluster A_i, where $i = 1,..., n$, and $j = 1,..., p$.

The condition "P is a fuzzy partition of X" implies:

$$\sum_{i=1}^{n} A_i(X^j) = 1,\ j = 1,..., p.$$

Let us denote by L^i the prototype of fuzzy class A_i, $i = 1,..., n$. We denote with $L = (L^1, L^2, ..., L^n)$ the representation of the fuzzy partition P.

A prototype of a cluster can be formed as a group of characteristics, each characteristic is a group of ordered pairs, each is of the form $(H_{qk}, e_{qk,i})$, where H_{qk} is the q-th distinct value of the k-th characteristic and $e_{qk,i}$ is the degree of association of this value to the k-th characteristic in the L^i prototype such that:

$$L_k^i = \left[\left(H_{1k}, e_{1k,i}\right), \left(H_{2k}, e_{2k,i}\right), \ldots, \left(H_{qk}, e_{qk,i}\right)\right], \qquad k = 1,\ldots, s,$$

where L_k^i is the k-th characteristic of the i-th prototype. We denote with X_k^j the value of the k-th characteristic for the object X^j and we have the following relations:

$$(3) \quad \cap_{l=1}^{q} H_{lk} = \varnothing, \ \bigcup_{l=1}^{q} H_{lk} = \bigcup_{j=1}^{p} X_k^j, \ 0 \le e_{lk,i} \le 1; \ \sum_{l=1}^{q} e_{lk,i} = 1,$$

where: q is the number of distinct values for the k-th characteristic; $k = 1,2,\ldots, s$; $j = 1,2,\ldots, p$; $i = 1,2,\ldots, n$.

We have $e_{lk,i} = 0$, if the value associated with it is not a part of the k-th characteristic, while $e_{lk,i} = 1$, if there are no distinct values sharing this value in forming the characteristic.

Example of a prototype for the objects described in Table I

A prototype for the objects described in Table I can be:

DISPLAY	RAM	ROM	MP	KEYS
(COLOR TV., 0.31)	(48, 0.62)	(10, 0.38)	(6502, 0.28)	(52, 0.08)
(B&W TV., 0.40)	(32, 0.18)	(11-16, 0.05)	(6502A, 0.07)	(57-63, 0.25)
(BUILT-IN, 0.24)	(64, 0.20)	(4, 0.14)	(Z80, 0.44)	(64-73, 0.36)
(TERMINAL, 0.15)		(1, 0.11)	(8080, 0.08)	(92, 0.05)
		(8, 0.12)	(HP, 0.05)	(53-56, 0.26)
		(80, 0.03)	(6502C, 0.08)	
		(12, 0.02)		
		(14, 0.15)		

Figure 1. Representation of a prototype.

We have five characteristics for each characteristic we have a set of distinct values, each of this value with a weight denoted with $e_{lk,i}$ in (3).

For the **DISPLAY** characteristic we have four distinct values which are: *COLOR TV.*, *B&W TV.*, *BUILT-IN* and *TERMINAL*. Each of this distinct value has a weight between zero and one and the sum of these weights is one. And, of course this thing is true for all the characteristics, as we see in Fig. 1. The others characteristics are: **RAM**, **ROM**, **MP** and **KEYS**.

Now we define the distance measure between a symbolic object and a prototype of a cluster. The dissimilarity between j-th object and i-th prototype is given by:

$$(4) \quad D\left(X^j, L^i\right) = \sum_{k=1}^{s} \sum_{l=1}^{q} D(X_k^j, H_{lk}) \cdot e_{lk,i}, \qquad j = 1,\ldots, p; \ \ i = 1,\ldots, n,$$

where s is the number of characteristics, q is the number of distinct values for the k-th characteristic and $D(X_k^j, H_{lk})$ is the dissimilarity between the value of the k-th characteristic in the j-th object and the l-th distinct value of k-th characteristic in the i-th prototype, and this dissimilarity is calculated using the formulas given in (1) and (2), depends on the k-th characteristic type.

Example of dissimilarity calculation between an object and a cluster center.

For this example we consider the cluster center described in Fig.1 and the object with number 3 from Table I.

The characteristic values for this object are:

Characteristic Name	Characteristic Value	Characteristic Type
DISPLAY	B&W TV.	Qualitative
RAM	48K	Quantitative
ROM	4K	Quantitative
MP	Z80	Qualitative
KEYS	57-63	Quantitative

According to (4) the dissimilarity between the prototype given in Figure 1 and the object "Exidi sorcerer" is:

D= D(B&W TV., COLOR TV.) ·0.31 + D(B&W TV., B&W TV.) ·0.40
+ D(B&W TV., BUILT-IN)·0.24 + D(B&W TV., TERMINAL)·0.15
+D(48K, 48K) · 0.62 + D(48K, 32K) · 0.18 + D(48K, 64K) · 0.20
+D(4K, 10K) · 0.38 + D(4K, 11-16K) · 0.05 + D(4K, 4K) · 0.14
+D(4K, 1K) · 0.11 + D(4K, 8K) · 0.12 + D(4K, 80K) · 0.03
+D(4K, 12K) · 0.02 + D(4K, 14K) · 0.15
+D(Z80, 6502) · 0.28 + D(Z80, 6502A) · 0.07 + D(Z80, Z80) · 0.44
+D(Z80, 8080) · 0.08 + D(Z80, HP) · 0.05 + D(Z80, 6502C) · 0.08
+D(57-63, 52) · 0.08 + D(57-63, 57-63) · 0.25 + D(57-63, 64-73) · 0.36
+D(57-63, 92) · 0.05 + D(57-63, 53-56) · 0.26

DISPLAY Characteristic
RAM Characteristic
ROM Characteristic
MP Characteristic
KEYS Characteristic

Consider the criterion function **J** defined as:

$$(5)\quad J(P,L)=\sum_{i=1}^{n}\sum_{j=1}^{p}\left(A_i(X^j)\right)^m\cdot D^2\left(X^j,L^i\right),$$

where m > 1.

We are lead to the optimization problem:

$$\begin{cases} \text{minimize } J(P,L) \\ \text{subject to :} \\ 0\le A_i(X^j)\le 1, \qquad j=1,\ldots,p; \quad i=1,\ldots,n; \\ \sum_{i=1}^{n} A_i(X^j)=1, \quad j=1\ldots,p; \\ 0<\sum_{j=1}^{p} A_i(X^j)<p, \qquad i=1,\ldots,n; \end{cases}$$

Using the Lagrange multiplier method we may compute the optimal fuzzy partition for a fixed representation L.

The optimal partition P may be call the fuzzy partition induced by the prototypes $L^1, L^2, ..., L^n$.

The optimization problem will be solved by using a Genetic Algorithm (see [5], [10] and [3]). The genetic algorithm approach will be presented in the next section.

4. Symbolic clustering using a Genetic Algorithm

4.1. Chromosomes representation

The representation of a chromosome is:

$$c = \left(L^1, L^2, ..., L^n\right),$$

where $L^1, L^2, ..., L^n$ are the prototypes. We denote by **r** the number of the chromosomes.

4.2. Fitness function

Let c be a chromosome, $c = (L) = (L^1, L^2, ..., L^n)$ and let P be the fuzzy partition introduced by L. The fitness value f(c) of c is defined as:

$$f(c) = Max - J(P, L),$$

where *Max* is a real value chosen in such way that f(c) ≥ 0, for each chromosome *c*. The *Max* value can be equal to maximum value of the function *f*. The values for dissimilarity between the object characteristics are normalized between zero and one. Therefore we can find a reasonable value that is greater than the maximum value of the function **J**.

4.3. The selection operator

The selection method that we apply is the proportional selection, using the Monte Carlo algorithm. For this we calculate the following values:

$$F = \sum_{i=1}^{r} f(c_i),$$

where $P(t) = \{c_1, c_2, ..., c_r\}$, represents the actual population at the t-th moment.

The selection probability p_i for the c_i chromosome is defined as:

$$p_i = \frac{f(c_i)}{F}, \qquad i = 1, ..., r.$$

4.4. The crossover operator

Let us consider two chromosomes selected for the crossover operation:

$$C = (C^1, C^2, ..., C^n), \; D = \left(D^1, D^2, ..., D^n\right)$$

C^i and D^i are prototypes of the i-th cluster, i=1,…,n.

Let $F = (F^1, F^2, ..., F^n)$ be the unique offspring. We may consider F is obtained using convex combination of the genes in C and D.

The i-th gene in C is:

$$C^i = (C_1^i, C_2^i, ..., C_s^i), \quad i = 1,...,n.$$

where s is the number of characteristics.

The k-th component of the i-th gene is:

$$C_k^i = \left[(H_{1k}, e_{1k,i}^c), (H_{2k}, e_{2k,i}^c), ..., (H_{qk}, e_{qk,i}^c) \right], \quad k = 1,...,s; \ i = 1,...,n,$$

where q is the number of distinct values for the k-th characteristic and the upper index c in the weight: $e_{lk,i}^c$, $l = 1,...,q$ means that this weight refers to the chromosome C.

The weights $e_{lk,i}^f$, $l = 1,...,q$ of the chromosome F are calculated as follows:

$$e_{lk,i}^f = \alpha \cdot e_{lk,i}^c + (1-\alpha) \cdot e_{lk,i}^d, \ l = 1,...,q; \ k = 1,...,s; \ i = 1,...,n,$$

where α is a random number having the uniform distribution on [0,1].

The conditions (3) may be expressed as:

$$\begin{cases} \sum_{l=1}^{q} e_{lk,i}^f = 1, \\ 0 \le e_{lk,i}^f \le 1, \end{cases}$$

where $i = 1,...,n$, $k = 1,...,s$, and q is the number of distinct values of the k-th characteristic. It is easy to prove that these conditions are fulfilled:

a). The sum can be decomposed as follows:

$$\sum_{l=1}^{q} e_{lk,i}^f = \sum_{l=1}^{q} \left[\alpha \cdot e_{lk,i}^c + (1-\alpha) \cdot e_{lk,i}^d \right] = \alpha \cdot \sum_{l=1}^{q} e_{lk,i}^c + (1-\alpha) \cdot \sum_{l=1}^{q} e_{lk,i}^d = \alpha + (1-\alpha) = 1.$$

b). The condition $e_{lk,i}^f \ge 0$ is fulfilled since:

$$e_{lk,i}^f = \alpha \cdot e_{lk,i}^c + (1-\alpha) \cdot e_{lk,i}^d \ge \alpha \cdot 0 + (1-\alpha) \cdot 0 = 0.$$

c). The condition $e_{lk,i}^f \le 1$ is fulfilled since:

$$e_{lk,i}^f = \alpha \cdot e_{lk,i}^c + (1-\alpha) \cdot e_{lk,i}^d \le \alpha \cdot 1 + (1-\alpha) \cdot 1 = 1.$$

Example: For the objects described in Table I, suppose we must apply the crossover operator for two chromosomes. Let us consider the first gene in the first chromosome is:

DISPLAY	RAM	ROM	MP	KEYS
(COLOR TV., 0.27)	(48, 0.5)	(10, 0.2)	(6502, 0.08)	(52, 0.03)
(B&W TV., 0.5)	(32, 0.22)	(11-16, 0.02)	(6502A, 0.17)	(57-63, 0.05)
(BUILT-IN, 0.11)	(64, 0.28)	(4, 0.07)	(Z80, 0.24)	(64-73, 0.46)
(TERMINAL, 0.12)		(1, 0.18)	(8080, 0.28)	(92, 0.35)
		(8, 0.1)	(HP, 0.12)	(53-56, 0.11)
		(80, 0.3)	(6502C, 0.11)	
		(12, 0.04)		
		(14, 0.09)		

The first gene of the second chromosome involved in the crossover is:

DISPLAY	RAM	ROM	MP	KEYS
(COLOR TV., 0.11)	(48, 0.12)	(10, 0.12)	(6502, 0.09)	(52, 0.27)
(B&W TV., 0.21)	(32, 0.48)	(11-16, 0.25)	(6502A, 0.24)	(57-63, 0.09)
(BUILT-IN, 0.34)	(64, 0.4)	(4, 0.03)	(Z80, 0.16)	(64-73, 0.45)
(TERMINAL, 0.34)		(1, 0.05)	(8080, 0.25)	(92, 0.05)
		(8, 0.23)	(HP, 0.11)	(53-56, 0.14))
		(80, 0.20)	(6502C, 0.15)	
		(12, 0.04)		
		(14, 0.08)		

The expression of the first gene of the offspring obtained for $\alpha = 0.3$ is:

DISPLAY	RAM	ROM	MP	KEYS
(COLOR TV., 0.158)	(48, 0.234)	(10, 0.144)	(6502, 0.087)	(52, 0.198)
(B&W TV., 0.297)	(32, 0.402)	(11-16, 0.181)	(6502A, 0.219)	(57-63, 0.078)
(BUILT-IN, 0.271)	(64, 0.364)	(4, 0.042)	(Z80, 0.184)	(64-73, 0.453)
(TERMINAL, 0.274)		(1, 0.089)	(8080, 0.259)	(92, 0.140)
		(8, 0.191)	(HP, 0.113)	(53-56, 0.131)
		(80, 0.230)	(6502C, 0.138)	
		(12, 0.040)		
		(14, 0.083)		

This gene is computed in the sequel:

The **DISPLAY** characteristic:

COLOR TV.: $\alpha \cdot 0.27 + (1- \alpha) \cdot 0.11 = 0.158$,

B&W TV.: $\alpha \cdot 0.50 + (1- \alpha) \cdot 0.21 = 0.297$,

BUILT-IN: $\alpha \cdot 0.11 + (1- \alpha) \cdot 0.34 = 0.271$,

TERMINAL: $\alpha \cdot 0.12 + (1- \alpha) \cdot 0.34 = 0.274$.

The **RAM** characteristic:

48K: $\alpha \cdot 0.50 + (1- \alpha) \cdot 0.12 = 0.234$,

32K: $\alpha \cdot 0.22 + (1- \alpha) \cdot 0.48 = 0.402$,

64K: $\alpha \cdot 0.28 + (1- \alpha) \cdot 0.40 = 0.364$.

The **ROM** characteristic:

10K: $\alpha \cdot 0.20 + (1- \alpha) \cdot 0.12 = 0.144$,

11-16K: $\alpha \cdot 0.02 + (1- \alpha) \cdot 0.25 = 0.181$,

4K: $\alpha \cdot 0.07 + (1- \alpha) \cdot 0.03 = 0.042$,

1K: $\alpha \cdot 0.18 + (1- \alpha) \cdot 0.05 = 0.089$,

8K: $\alpha \cdot 0.10 + (1- \alpha) \cdot 0.23 = 0.191$,

80K: $\alpha \cdot 0.30 + (1- \alpha) \cdot 0.20 = 0.230$,

12K: $\alpha \cdot 0.04 + (1- \alpha) \cdot 0.04 = 0.040$,

14K: $\alpha \cdot 0.09 + (1- \alpha) \cdot 0.08 = 0.083$.

The **MP** characteristic:

6502: $\alpha \cdot 0.08 + (1- \alpha) \cdot 0.09 = 0.087$,

6502A: $\alpha \cdot 0.17 + (1- \alpha) \cdot 0.24 = 0.219$,

Z80: $\alpha \cdot 0.24 + (1-\alpha) \cdot 0.16 = 0.184$,
8080A: $\alpha \cdot 0.28 + (1-\alpha) \cdot 0.25 = 0.259$,
HP: $\alpha \cdot 0.12 + (1-\alpha) \cdot 0.11 = 0.113$,
6502C: $\alpha \cdot 0.11 + (1-\alpha) \cdot 0.15 = 0.138$.

The **KEYS** characteristic:

52: $\alpha \cdot 0.03 + (1-\alpha) \cdot 0.27 = 0.198$,
57-63: $\alpha \cdot 0.05 + (1-\alpha) \cdot 0.09 = 0.078$,
64-73: $\alpha \cdot 0.46 + (1-\alpha) \cdot 0.45 = 0.453$,
92: $\alpha \cdot 0.35 + (1-\alpha) \cdot 0.05 = 0.140$,
53-56: $\alpha \cdot 0.11 + (1-\alpha) \cdot 0.14 = 0.131$.

4.5. The mutation operator

Assume we have selected a gene for the mutation operation. Let us denote with L^i this gene. The mutated gene denoted with $L^{'i}$, is obtained using the one of the following methods:

Method 1

In this first method we choose randomly a component L_k^i of the L^i gene and modify the value for this component only.

1. $L^i = \left(L_1^i, L_2^i, \ldots, L_s^i\right), \quad i \in \{1, \ldots, n\};$
   $L^{'i} = \left(L_1^i, L_2^i, \ldots, L_k^{'i}, \ldots, L_s^i\right), \quad i \in \{1, \ldots, n\};$
   where **s** is the number of characteristics.
2. Chose a random value k, $k \in \{1, \ldots, s\}$.
3. Let us consider the component L_k^i of the L^i gene.
$$L_k^i = \left[\left(H_{1k}, e_{1k,i}\right), \left(H_{2k}, e_{2k,i}\right), \ldots, \left(H_{qk}, e_{qk,i}\right)\right]$$
The mutated component will be:
$$L_k^{'i} = \left[\left(H_{1k}, e'_{1k,i}\right), \left(H_{2k}, e'_{2k,i}\right), \ldots, \left(H_{qk}, e'_{qk,i}\right)\right]$$
4. For each value $e_{lk,i}$, l = 1,...,q, generate a random number y, following the normal distribution N(0,σ), where σ is a value chosen in a convenient way.
   Calculate $e'_{lk,i} = e_{lk,i} + y, \quad l = 1, \ldots, q.$
5. Compute: $S = \sum_{l=1}^{q} e'_{lk,i}$.

Normalize the weights according to: $e'_{lk,i} = \dfrac{e'_{lk,i}}{S}$.

Method 2

In this second method we modify the values for all components of the selected gene.

1. $L^i = \left(L_1^i, L_2^i, \ldots, L_s^i\right), \quad i \in \{1, \ldots, n\};$

$L^{'i} = \left(L_1^{'i}, L_2^{'i}, ..., L_s^{'i}\right), \quad i \in \{1,...,n\};$

where **s** is the number of characteristics.

2. Let us consider that the components of the L^i gene are:

$$L_k^i = \left[\left(H_{1k}, e_{1k,i}\right), \left(H_{2k}, e_{2k,i}\right), ..., \left(H_{qk}, e_{qk,i}\right)\right],$$

The mutated components will be:

$$L_k^{'i} = \left[\left(H_{1k}, e'_{1k,i}\right), \left(H_{2k}, e'_{2k,i}\right), ..., \left(H_{qk}, e'_{qk,l}\right)\right],$$

3. For each value $e_{lk,l}$, l = 1,...,q, generate a random number y, following the normal distribution N(0,σ), where σ is a value chosen in a convenient way.

Calculate $e'_{lk,i} = e_{lk,i} + y, \quad l = 1,...,q$.

4. Compute: $S = \sum_{l=1}^{q} e'_{lk,i}$.

Normalize weights according to: $e'_{lk,i} = \dfrac{e'_{lk,i}}{S}$.

Example: This example considers the objects described in Table I. Suppose we have a chromosome selected for mutation and in this chromosome we have the first gene selected for the mutation.

If we apply the first method and suppose that the value of k is equal to 1 we obtain that the characteristic used for mutation is **DISPLAY** and q = 4 (number of distinct values for this characteristic), so we have four random values for **y**.

Suppose we have the initial values for this characteristic:

DISPLAY
(COLOR TV., 0.33)
(B&W TV., 0.14)
(BUILT-IN, 0.24)
(TERMINAL, 0.29)

and we have the following generated values for **y**:

COLOR TV.: For **y** = 3.1 we have

$$e'_{11,1} = e_{11,1} + y = 0.33 + 3.1 = 3.43.$$

B&W TV.: For **y** = 1.2 we have

$$e'_{21,1} = e_{21,1} + y = 0.14 + 1.2 = 1.34.$$

BUILT-IN: For **y** = 2.3 we have

$$e'_{31,1} = e_{31,1} + y = 0.24 + 2.3 = 2.54.$$

TERMINAL: For **y** = 4.4 we have

$$e'_{41,1} = e_{41,1} + y = 0.29 + 5.4 = 4.69.$$

The sum of these values is:

$$S = \sum_{l=1}^{4} e'_{l1,1} = 3.43 + 1.34 + 2.54 + 5.69 = 12.$$

The new values $e'_{lk,1}$, $k = 1$, $l = 1,...,4$ are:

$$e'_{11,1} = \frac{e'_{11,1}}{S} = \frac{3.43}{12} = 0.286, \qquad e'_{21,1} = \frac{e'_{21,1}}{S} = \frac{1.34}{12} = 0.112,$$

$$e'_{31,1} = \frac{e'_{31,1}}{S} = \frac{2.54}{12} = 0.212, \quad e'_{41,1} = \frac{e'_{41,1}}{S} = \frac{4.69}{12} = 0.390.$$

Therefore the new weights for the characteristic DISPLAY are:

DISPLAY
(COLOR TV., 0.286)
(B&W TV., 0.112)
(BUILT-IN, 0.212)
(TERMINAL, 0.390)

With the second method we compute all five characteristics in the same way as we compute the characteristic **DISPLAY**.

4.6. The initial population

For each characteristic of a prototype we must obey the rules described in [3]. Therefore for each characteristic of each prototype in each chromosome we may choose randomly a weight $e_{lk,i}$, that will be equal to 1, and the remainder terms in that characteristic will have the value 0.

4.7. The stop condition

The stop condition for the algorithm is to specify a maximum number of generations N, and at the final we retain the best individual of all generations.

4.8. The fuzzy partition

The formula that we used to calculate the fuzzy partition is given in [1], [8], in the Fuzzy Symbolic C-Means Algorithm (FSCM):

$$A_i(X^j) = \begin{cases} \dfrac{1}{\sum_{q=1}^{n} \left(\dfrac{D(X^j, L^i)}{D(X^j, L^q)} \right)^{\frac{2}{m-1}}}, & L^i \neq X^j, \\ 1, & L^i = X^j. \end{cases}$$

The prototype L^i is equal to the object X^j $(L^i = X^j)$, if the degree of association $e_{lk,i}$ of the prototype L^i is equal to 1 for the characteristics values presents

in the object X^j and 0 for the remainder characteristics values that are not present in the object X^j.

5. The genetic algorithm

The genetic algorithm for symbolic clustering may be described as follows:

The Genetic Algorithm for Symbolic Clustering:

P1. Initialize the population P(0) according to **4.6** ; Set t := 0. (t is the number the generations).

P2. For each chromosome in the population P(t) compute the fuzzy partition using the formula given in Section **4.8.**

P3. The population P(t) are evaluated using the fitness function *f* described in Section **4.2.** The best individual in generation P(0) is recorded.

P4.

While (non C) execute: (condition C is described in Section **4.7**)

P4.1. From the population P(t) we select the chromosomes that will be used to obtain the new generation. We denote with P^1 the intermediate population. The selection method is given in Section **4.3.**

P4.2. For the chromosomes in P^1 we apply the genetic operators crossover and mutation which are described in Sections **4.4** and **4.5.** We denote by P^2 the new population. From the population P^1 we delete the parents of the chromosomes that are in P^2. The rest of the chromosomes that remain in P^1 are added to the P^2. The new generation is $P(t+1) := P^2$.

P4.3. Let t := t + 1; for each chromosome in the population P(t) compute the fuzzy partition according to **4.8.** Evaluate the population P(t) and record the best individual.

End While

REFERENCES

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum Press, 1981.

[2] E. Diday, *The Symbolic Approach in Clustering, Classification and Related Methods of Data Analysis*, H.H. Bock, Ed. Elsevier, Amsterdam 1988.

[3] D. Dumitrescu, *Algoritmi Genetici*, Ed. Albastra, Cluj-Napoca 1999.

[4] D. Dumitrescu, *Teoria Clasificarii*, Univ. "Babes-Bolyai", Cluj-Napoca, 1991.

[5] D.E. Goldberg, *Genetic Algorithms in Search Optimizations and Machine Learning*, Addison Wesley, Reading, MA., 1996.

[6] K.C. Gowda and E. Diday, *Symbolic clustering using a new dissimilarity measure*, Pattern Recogn., 24, 6, pp. 567-578, 1991.

[7] K.C. Gowda, E. Diday, *Symbolic clustering using a new similarity measure*, IEEE Trans. Syst., Man. Cybern., 22, pp. 368-378, 1992.

[8] Yasser El-Sonbaty, M.A. Ismail, *Fuzzy Clustering For Symbolic Data*, IEEE Transactions on Fuzzy Systems, 6, 2, 1998, pp. 195-204.

[9] R. Michalski, R.E. Stepp, *Automated construction of classifications: Conceptual clustering versus numerical taxonomy*, 5, 396-410, 1983.

[10] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA, 1994.

[11] E. R. Ruspini, *Numerical methods for fuzzy clustering*, Inform. Sci., 2, 318-350, 1970.

Babeş-Bolyai University, Faculty of Mathematics and Informatics, RO 3400 Cluj-Napoca, str. M. Kogalniceanu 1, România.

E-mail address: ddumitr@cs.ubbcluj.ro

Dipartimento di Ingegneria della Informazione, Universita di Pisa, Italia

E-mail address: beatrice@iet.unipi.it