EVOLUTIONARY HIERARCHICAL CLUSTERING TECHNIQUE

by
Monica Chiş

Abstract. In this paper an evolutionary technique for detecting hierarchical structure of a data set is considered [4].
A linear representation of the cluster structure within the data set is used. An evolutionary algorithm evolves a population of clustering hierarchies. Proposed algorithm uses mutation and crossover as search (variation) operators. Binary tournament selection is considered. A new crossover operator is used.
The number of clustering levels and the number of clusters on each level are subject to optimization.

1. INTRODUCTION

Clustering is an important technique used in the simplification of data sets or in discovering some inherent structure present in data. The purpose of cluster analysis is to partition a given data set into a number of groups such that data in a particular cluster are more similar to each other than objects in different clusters ([2]).
Hierarchical clustering groups data into a tree structure, thus building multi-level representation capable of revealing inter-cluster relationships.
Hierarchical clustering assumes a significant role in a variety of research areas such as data mining, pattern recognition, economics, biology, etc.
One of the issues in evolutionary algorithms (EAs) is the relative importance of two search operators. Genetic algorithms (GAs) and genetic programming stress the role of crossover while evolutionary progamming end evolution strategies stress the role of mutation [9].

In this paper an evolutionary technique for detecting hierarchical clustering structures is proposed. We use a technique in which mutation and crossover operators are combining. The representation we used is linear and may describe any tree structures. This representation is not limited to binary trees.
An evolutionary algorithm with mutation and crossover is used. Binary tournament selection is considered.

2. SOLUTION REPRESENTATION

Let $X=\{x^1, x^2,..., x^p\}$ be a data set and d a distance on X. For applying an evolutionary algorithm to solve hierarchical classification problem we have to build a hierarchical representation scheme. Each candidate solution (individual) describes a clustering hierarchy.

In this approach each individual (candidate solution or chromosome) describes a hierarchy. A linear sequence is used for representing the cluster hierarchy. The sequence is translated into a classification tree. Each node of the tree represents a cluster (a class). Evolved solutions represent trees of variable shape and size. Therefore the number of clustering levels and the number of clusters on each level are subject to optimization.

Classes of the classification tree are labelled by parsing each level from left to right and top to bottom. Root node is considered as a class containing the entire data set X. The root node has label 0. Clasification data are assigned to the terminal classes only.

Let p be the cardinality of data set X and $M(p)$ the maximum number of classes needed for representing the data set X by a clasification tree [4]

$$M(p) = 2(p - 1).$$

Proposed representation indicates how data points are assigned to classes (tree nodes) and how classes are related to another classes in the tree representation.
We consider chromosomes (individuals) are represented by linear structures of constant length[4].
Chromosome length L is given by

$$L = p + M(p).$$

An individual c is thus a sequence with two sections: a point-section and a class-section. Point-section of the chromosome c is representing by the first p symbols of c. Class-section of the chromosome c is representing by the last $M(p)$ symbols of c [4].

An individual c is a sequence:

$$c = (c_1, c_2, ..., c_p, c_{p+1}, ..., c_L),$$

where c_j is an integer number.
The first p symbols of c have to satisfy the requirement [4]:

$$1 \le c_j \le M(p),\ 1 \le j \le p\ .$$

Value c_j of the gene j, $1 \le j \le p$, indicates the terminal tree node (class) to which the point x^j is attached.

For symbols c_{p+1}, …, $c_{p+M(p)}$ the inequality

$$0 \le c_{p+j} \le j-1,\ j = 1, \ldots, M(p).$$

holds.

Value c_{p+j} indicates that A_j is a subclass of the class A c_{p+j}. If

$$c_{p+j} = 0$$

then the class A_j is a subclass of root node X.

Example 1

Consider data set $X=\{x^1, x^2, x^3, x^4, x^5, x^6\}$. In this case the maximum number of classes is

$$M(6) = 10.$$

The chromosome length is:

$$L\ = 6 + 8.$$

Consider the individual

$$c_1 = (4,5,6,2,7,7,0,0,0,1,1,3,3,3,3,0).$$

Representation of the individual c_1 is the tree depicted in Figure 1:

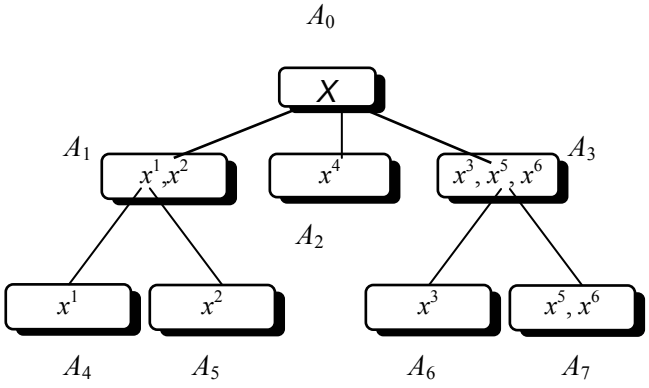

Figure 1. Tree associated to the individual $c_1 = (4,5,6,2,7,7,\ 0,0,0,1,1,3,3,3,3,\ 0)$.

Classes A_2, A_4, A_5, A_6, A_7 correspond to the terminal nodes. Thus we have:
$A_2 = \{x^4\}$,
$A_4 = \{x^1\}$,
$A_5 = \{x^2\}$,
$A_6 = \{x^3\}$,
$A_7 = \{x^5, x^6\}$,

Class A_1, A_3 correspond to the non-terminal nodes. In this case we have
$A_1 = \{x^1, x^2\}$,
$A_3 = \{x^3, x^5, x^6\}$.
Classes A_8, A_9, A_{10} are empty and for this reason it is not represented in our hierarchy.
The root node corresponds to $A_0 = X$.

Example 2

Consider data set $X = \{x^1, x^2, x^3, x^4, x^5\}$. In this case the maximum number of classes is

$$M(5) = 8.$$

The chromosome length is:

$$L = 5 + 8.$$

Consider the individual

$$c_2 = (1,1,3,2,3,0,0,0,0,0,0,0,0).$$

This individual describes the hierarchy depicted in Figure 2:

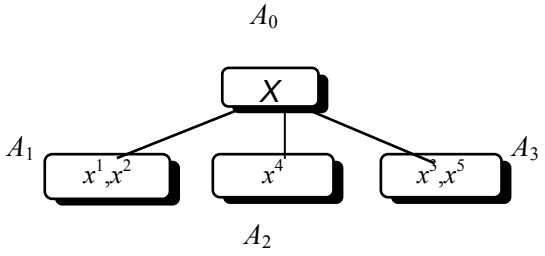

Figure 2. Tree associated to the individual $c_2 = (1,1,3,2,3,0,0,0,0,0,0,0,0)$.

Classes A_1, A_2, A_3 correspond to the terminal nodes. Thus we have:
$A_1 = \{x^1, x^2\}$
$A_2 = \{x^4\}$,
$A_3 = \{x^3, x^5\}$.
Classes A_4, A_5, A_6, A_7, A_8, are empty.

3. FITNESS FUNCTION

For detecting the hierarchy that represents the best cluster structure of X we need a fitness function to measure the hierarchy quality. Fitness function must have the best value for the hierarchy that is the most representative for input data set.

Our clustering criterion is based around minimising the sum of squared distances of objects from their cluster (class) prototypes divided by the cardinality of each cluster and for this purpose Euclidean distance is used. The optimal clusters hierarchy of data set X corresponds to the minimal fitness function value.

The prototype of a class A_i is given by:

$$L^i = \frac{\sum_{x^k \in A_i} x^k}{p_i},$$

where p_i is the cardinality of A_i.

We consider the fitness function:

$$eval(c) = \sum_{i=1}^{n} \sum_{x^k \in A_i} \frac{d^2(L_i, x^k)}{p_i}.$$

Fitness function is to be minimised.

4. EVOLUTIONARY ALGORITHM

A generational evolutionary algorithm is used to evolve a population of clustering hierarchies.

Proposed evolutionary algorithm uses mutation and crossover as search (variation) operators. Each gene of the individual in the population can be mutated with a given probability.
A crossover operator is consider and the crossover operator is combined with mutation.
We consider one point crossover. The crossover point is

$$k = p.$$

As a result of the crossover with $k = p$ two descendents are created. One of them have the class section of a parent and the other the class section of the other parent.
After crosover a mutation operator is used. Mutation is applied to each gene in the point section which not respect the condition

$$c_i \neq c_j; \forall\ 1 \leq i \leq p \text{ and } p+1 \leq j \leq M(p)+p$$

Mutation of the genes in point-section of the chromosome is restricted to indicating terminal classes only. This restriction ensures that no chromosome reparation is required [4].
If the mutated gene belongs to class-section of the chromosome it is possible that an invalid solution occurs. Therefore reparation is needed.
Let us consider the gene i is mutated in gene j. In this case the positions in point-section pointing towards A_i (having value i) have to be modified. For instance these positions may be assigned to the class A_i.

Binary tournament selection is considered.

Proposed Hierarchical Clustering Evolutionary Algorithm 3 (HCEA3) is outlined below:

Hierarchical Clustering Evolutionary Algorithm 3 (HCEA3)

begin
Set $t = 0$.
Initialize the population $P(t)$. {Random initialization is the basic choice}
while (**C**) **do**
Apply selection for $P(t)$. Let P^l be the set of the selected solutions.
The best 50 % individuals from current population are copied into $P(t+1)$.
All individual from $P(t)$ compete for solution.

Individuals from P^1 entering the mating pool based on tournament selection. Choose chromosomes from P^1 to enter the mating pool.
Apply the crossover operator to the solutions from the mating pool. A new intermediate population P^2 is obtained.
Mutate solutions in P^2 offspring enter the next generation $P(t+1)$.
Set $t = t + 1$.

end while

end

The condition **C** concerns the specified number of generations.
The problem solution is the best individual (hierarchy) obtained during the search process.

5. SUMMARY

Several data sets have been used for numerical experiments and the obtained results are encouraging. For Iris data set ([8]) the accuracy of the hierarchy is very good.
Solution accuracy generally increases with the number of generation. Usually, detected structures are not binary hierarchies.

Further research will explore another data sets and other operators in order to provide the ability to classify large data collections.

REFERENCES

[1]. Bäck T., Fogel D.B., Michalewicz, Z, Handbook of Evolutionary Computation, *Oxford University Press*, Oxford, 1997.
[2]. Bhuyan, J.N, Raghavan, V.V, Elayavalli, V.K, Genetic Algorithm for clustering with an ordered representation, *Proceedings of the Fourth International Conference on Genetic Algorithms and their Application*, 1991, 408-415.
[3]. Chiş, M., A new evolutionary hierarchical clustering technique, *Babeş-Bolyai University Research Seminars, Seminar on Computer Science*, 2000,13-20.
[4]. Dumitrescu D., Chiş, M., Evolutionary hierarchical clustering for data mining, *Babeş-Bolyai University, Zilele Academice Clujene, 17 iunie* 2002, (to appear).
[5]. Dumitrescu, D., Mathematical Principles of Classification (romanian), Romanian Academy Publishing House (Ed. Academiei Române), Bucureşti, 1999.

[6]. Dumitrescu, D., Lazzerini, B., Jain, L., Dumitrescu, A., Evolutionary Computation, C.R.C. Press, Boca Raton, FL., 2000.
[7]. Dumitrescu, D., Lazzerini, B., Hui, L., Hierarchical data structure detection using evolutionary algorithms, *Studia Babeş-Bolyai University, Ser.Informatica*, 2000.
[8]. Fisher, R. A., The use of multiple measurements in taxonomic problems, *Annals of Eugenics* 7, 1936, 179-188.
[9]. Spears W. M., Adapting Crossover in Evolutionary Algorithms, *Proceedings of the Evolutionary Programming Conference*, 1995,367-384.

Author:

Monica Chiş
Faculty of Applied Sciences, Avram-Iancu University, Cluj-Napoca,
PhD student at Babeş-Bolyai University, Cluj-Napoca,Faculty of Mathematics and Computer Science, Department of Computer Science.
mchis@personal.ro