Link-Cell Methods for Dynamic Evolutionary Clustering

Dan Dumitrescu[1], Ferenc Járai-Szabó[2], and Károly Simon[1]

[1] Babeş-Bolyai University,
Faculty of Mathematics and Computer Science, Computer Science Department,
{ddumitr,ksimon}@cs.ubbcluj.ro
[2] Faculty of Physics, Department of Biomedical Physics, jferenc@phys.ubbcluj.ro

Abstract. Evolutionary Algorithms can be successfully used for solving dynamic clustering problems. Based on a new evolutionary multi-modal optimization metaheuristics - called Genetic Chromodynamics (GC) - recently a new dynamic clustering algorithm has been proposed. Inspired from Computational Physics a Link-Cell-based method is proposed in order to obtain an improved GC model. The proposed Link-Cell technique is combined with the GC-based dynamic clustering algorithm (GCDC). In this way a new method for dynamic cluster detection (LCGCDC) is developed. Some numerical experiments are described.

AMS Subject Classification: 68W01
Keywords and phrases: evolutionary algorithm, cluster detection

1 Introduction

Evolutionary algorithms are useful tools for solving complex optimization search problems (see [3]). Many real world problems allow multiple solutions, which may be optimal or almost optimal. In order to identify several optima, special evolutionary models have been proposed (see [2,3,9]). New evolutionary algorithms for solving multi-modal optimization problems have been developed [4]. These algorithms generally use sub-population models and are based on a local interaction principle.

In order to promote local search and to obtain a general technique for improving evolutionary multi-modal optimization algorithms, a *Link-Cell* method is proposed. The main idea of this method is inspired from Computational Physics [12]. Link-cell methods are used especially in molecular dynamics simulations in order to handle short-range interactions and in this way to reduce the complexity of algorithms (see [1,12]). The fundamental idea is that the simulation cell is partitioned into a number of smaller sub-cells. At each time-step a linked list of all particles contained in each sub-cell is constructed. In this way short-range interactions between particles can be easily calculated taking into account only the particles which are in one sub-cell and in its first-order neighborhood.

Recently a new evolutionary search and multi-modal optimization metaheuristics - called Genetic Chromodynamics (GC) [4] - has been proposed. This

metaheuristics was used to derive new evolutionary algorithms for multi-modal optimization. Based on GC metaheuristics a new evolutionary dynamic clustering method has been proposed [5]. This GC-based dynamic clustering technique - called GCDC - has been successfully used for solving some practical problems (see [5,6]).

Dynamic clustering is a typical multi-modal optimization problem. By clustering a data set is divided into regions of high similarity, as defined by a distance metric. In most instances, a prototypical vector identifies a cluster. Hence, the problem of cluster optimization is twofold: optimization of cluster centers and determination of number of clusters. The latter aspect has often been neglected in standard approaches (static clustering methods) (see [10,11]), as these typically fix the number of clusters *a priori*. In case of practical problems the number of existing clusters is generally unknown. Opposed to static, dynamic clustering does not require *a priori* specification of the number of clusters. Evolutionary multi-modal optimization models proved to be useful tools for deriving dynamic clustering algorithms.

Proposed Link-Cell method is combined with the GC metaheuristics, for promoting GC local search. The obtained model is applied to GCDC algorithm. The Link-Cell technique is used for improving GCDC. The convergence of standard GCDC technique is influenced by some parameters. Some adaptation mechanisms for these parameters and some tentative to improve GCDC are known (see [7,8]). Using the proposed Link-Cell technique new parameter adaptation algorithms are derived for achieving a better performance. A new evolutionary algorithm for dynamic clustering - called LCGCDC - is obtained.

In Section 2 Link-Cell method is presented. There is described, how this method could be used for obtaining improved evolutionary multi-modal optimization models. The method is combined with the GC metaheuristics. The standard GC-based clustering algorithm is presented in Section 3. The new Link-Cell-based GCDC method is described in Section 4. Section 5 presents numerical experiments for performance investigation of the new clustering algorithm.

2 Link-Cell-Based Evolutionary Multi-Modal Optimization

In certain situations we are interested not only in finding the global optima of a problem, but also in identifying the set of all acceptable solutions. Typical cases are multi-modal function optimizations, covering and clustering problems.

Standard evolutionary algorithms quickly concentrate the search effort in the most promising regions of search space. These algorithms tend to converge to a single solution, to the global optima of the problem. In order to identify more optimum points, the evolutionary algorithm has to be endowed with additional mechanisms aiming to favor and preserve population diversity. This may be accomplished by promoting local search and allowing evolutionary algorithms to evolve sub-populations. Special evolutionary models for realizing these goals are known (see [2,3,9]).

Standard evolutionary multi-modal optimization models, like niching techniques, in some particular situations cannot focus the search on each optimum and find the optimal solutions efficiently. Some of the search effort is wasted in recombination of inter-optimum solutions. Combining two solutions from different sub-populations may produce lethal solutions. Identification of the number of optimal solutions could be a problem, as well.

Recently a new evolutionary metaheuristics - called Genetic Chromodynamics (GC) [4] - has been proposed to overcome the shortcomings of classical evolutionary multi-modal optimization models.

Genetic Chromodynamics Genetic Chromodynamics is a new kind of evolutionary search and multi-modal optimization metaheuristics. GC-based methods use a variable-sized population, a stepping-stone search mechanism, a local interaction principle and a new operator for merging very close individuals.

Corresponding to stepping-stone technique each individual in the population has the possibility to contribute to the next generation and thus to search progress. According to local interaction principle only short-range interactions between solutions are allowed.

To enhance GC, micropopulation models [4] can be used. Corresponding to these models, for each individual a local interaction domain is considered. Individuals within this domain represent a micropopulation. All solutions from a micropopulation are recombined. When the local domain of an individual is empty the individual is mutated.

Within GC sub-populations co-evolve and eventually converge toward several optima. The number of individuals in current population usually changes with the generation. A merging operator is used for merging very close individuals. At convergence, the number of sub-populations equals the number of optima. Each final sub-population hopefully contains a single individual representing an optima, a solution of the problem.

Link-Cell-Based Local Search for Evolutionary Multi-Modal Optimization Inspired from Computational Physics a Link-Cell-based model is proposed to obtain a general technique for improving evolutionary multi-modal optimization algorithms.

According to Link-Cell technique the search space is partitioned into smaller, interconnected sub-cells. At each generation (time-step) a list of all chromosomes (particles) contained in each cell is constructed.

This Link-Cell model could be efficient for handling short-range interactions between individuals. For a chromosome the k-th order neighborhood of the cell containing the chromosome can be considered as interaction domain. In this way short-range interactions between solutions can be easily calculated.

An adaptation mechanism for controlling the size of interaction domain (the value of parameter k) can be used. Sub-population stabilization can be promoted by adapting individuals interaction domains.

Evolutionary algorithms are generally influenced by some method parameters. The Link-Cell technique could be efficient for deriving some new parameter adaptation techniques.

The proposed Link-Cell method can be combined with any evolutionary multi-modal optimization model. Combining Link-Cell technique with GC a new model for evolutionary multi-modal optimization is obtained. This new model can be used for improving GC-based algorithms.

3 Standard Genetic Chromodynamics-Based Dynamic Clustering

Based on GC metaheuristics recently a new dynamic clustering algorithm has been proposed. The standard GCDC algorithm is described below.

Solution Representation and Fitness Assignment Each cluster is represented by a prototype (cluster center). Each prototype is encoded into a chromosome.

The idea of GCDC method is to determine sub-populations of evolving chromosomes converging toward prototypes of real clusters.

The initial population is randomly generated and it contains a large number of individuals. Fitness values of individuals are evaluated using suitable fitness functions. For instance Gaussian functions can be used (see [7, 8]).

Interaction Domain For realizing the local interaction principle, an interaction domain (mating region) is considered for each individual in the population (a chromosome representing a prototype).

To support sub-population stabilization an adaptation mechanism can be used for controlling interaction domains (see [8]). Within this adaptation mechanism the interaction range of each individual could be different.

Population Model For realizing the stepping-stone search principle at each step of the generation process each chromosome is selected to produce an offspring through crossover or mutation.

A micropopulation model is used. The crossover mate for an individual (dominant parent) is selected among the chromosomes in its interaction domain. Only one offspring is generated. If there is no mate in the interaction domain of an individual, then the mutation operator will be applied.

An offspring can replace only its dominant parent. The most fitted between dominant parent and offspring is introduced in the new generation.

An effect of crossover operation is that chromosomes in the same sub-population partially overlap after a certain number of iterations. When the distance between two chromosomes is smaller than a considered value ε (*merging radius*) the chromosomes are merged. In this way the size of the population decreases during the search process. Final population contains as many individuals as the optimal cluster number.

Search Operators Within GCDC any type of known search operators can be used. For instance the crossover operation can be a convex combination of the parent genes. A randomly generated number for each gene can be considered as combination coefficient.

An additive perturbation of genes with a randomly chosen value from a normal distribution N(0,σ), where σ is a control parameter called *mutation step size* can be considered as mutation operator.

Termination If no more changes occur in the population through a fixed number of iterations, then the search process stops. The individuals within last population are considered as prototypes of naturally existing clusters.

4 Link-Cell-Based GCDC

The proposed Link-Cell method (see Section 2) is applied to GCDC. The method is used for improving GCDC by promoting local search and deriving new parameter adaptation techniques.

Link-Cell Division According to the Link-Cell technique the search space is partitioned into small interconnected cells. In the proposed clustering technique the dimension of a cell is computed using the minimum distance between input samples (data points).

The initial population is randomly generated. In addition a chromosome for each cell can be considered for achieving a better exploration.

For a chromosome the k-th order neighborhood of the cell containing the chromosome is considered as interaction domain. An adaptation technique is used for controlling interaction domains.

Interaction Domain Adaptation Sub-population stabilization can be promoted by adapting individuals interaction domains. At last stages of search process the dimension of the interaction domain of a chromosome would be close to the diameter of corresponding cluster. For achieving this goal, at beginning a small interaction domain is considered for each prototype. This interaction domain is extended during the search progress.

Initially, as interaction range the first-order neighborhood of the prototype is considered. At each generation interaction domains of all individuals are recalculated using the algorithm described below.

For a chromosome L_j the current interaction domain D_j is evaluated. Let N_j be the set of points in D_j. The next-order neighborhood of the individual is considered as an extended interaction domain D_j^*. Let N_j^* be the set of points in D_j^*.

If the interaction domain of the individual is empty ($N_j = \emptyset$), then it will be extended. As new interaction domain the next-order neighborhood D_j^* will be considered. If there are data points in D_j, then the set N_j^* will be evaluated. If

$N_j^* \backslash N_j$ is empty, then the previous interaction domain D_j will not be modified. Else, the extended interaction domain D_j^* will be considered as new interaction domain for the individual.

The interaction domain adaptation (IDA) algorithm can be described as follows:

begin IDA for chromosome L_j
Calculate N_j (the set of points in D_j);
Calculate N_j^* (the set of points in D_j^*);
If $(N_j = \emptyset)$ then $D_j := D_j^*$;
else if $(N_j^* \backslash N_j \neq \emptyset)$ then $D_j := D_j^*$;
end IDA for chromosome L_j

Dynamical Fitness Function The set of input samples $X = \{x_1, ..., x_n\}$ is considered. Cluster structure corresponding to this input data set is given by a set of prototypes $L = \{L_1, ..., L_m\}$, represented by chromosomes. Fitness of a chromosome L_j is calculated using the following Gaussian fitness function:

$$g(L_j) = \sum_{i=1}^{m} e^{-\frac{\|x_i - L_j\|^2}{\gamma_j^2}}. \quad (1)$$

Parameters of corresponding normal distribution are L_j and γ_j.

An adaptation mechanism is used for controlling parameter $\gamma_j, j = 1, ..., m$. Variance parameter γ_j is recomputed for each individual L_j .

A dependence between variance γ_j and the interaction domain of L_j is introduced. In this way a dynamical adaptation of the fitness function is realized. At each generation the variance corresponding to the new interaction domain is recomputed. For each prototype L_j the parameter γ_j is computed using the mean distance between the points in interaction domain D_j.

If the interaction domain of L_j is empty, then the variance is computed using the diameter of the domain D_j.

Using this dynamical fitness assignment technique a more accurate detection of cluster prototypes is expected.

Setting and Adapting Parameters Variation operators described in Section 3 are used. Adaptation mechanism for controlling mutation step is considered. A method for setting merging distance parameter is also proposed.

The mutation step size is computed using the diameter of the interaction domain. Corresponding to this technique a mutated offspring belongs to the interaction domain of its parent. This mechanism could be efficient for preventing optima extinction.

Two chromosomes will be merged if they are within the same cell. In this way, cell dimension may also serve for computing merging distance.

After a final merging (see Section 4.5) individuals in the last population are considered as prototypes for the naturally existing clusters. Using obtained prototypes the cluster membership is computed for all data samples.

Final Merging According to the standard procedure if two chromosomes are in the same cell, then they are merged. This merging condition seems to be too strong for some particular situations. In some cases the distance between two chromosomes may be very small in spite of they belong to different cells. For preventing this drawback a final merging mechanism can be performed on the last population.

Final merging is based on the following rules:

- if the interaction domains of two different individuals contain the same sample points, then these individuals will be merged;
- two individuals will be merged, if each of them is in the interaction domain of the another.

Post-Processing and Fine Tuning Several numerical experiments revealed, in some particular cases, a small difference between the number of naturally existing clusters and the number of clusters detected by GCDC technique.

For instance, in some cases three prototypes for two real clusters. This could happen when there is only a small distance between these clusters. Lethal solutions (empty clusters) could also remain in final population.

To overcome these shortcomings a post-processing technique is proposed. This technique is based on the following rules:

(i) if a cluster can be expressed as the union of other clusters the chromosome representing the prototype of this cluster will be eliminated from final population;
(ii) the prototype of an empty cluster is eliminated from final population.

After steps (i) and (ii) a fine tuning mechanism could be performed for moving each detected prototype toward the mass center of the corresponding cluster. In this way the naturally existing cluster structure can be detected.

5 Numerical Experiments

Two numerical experiments concerning the use of LCGCDC are described. In first experiment the convergence of LCGCDC is investigated. In second experiment LCGCDC technique is compared with standard GCDC method.

Experiment 1. Convergence of the Link-Cell-Based GCDC Method Consider the data set $X = \{x_1, x_2, ..., x_{19}\}$, $x_i \in [100, 300] \times [100, 300]$. The data set X and the corresponding link-cell division are depicted in Figure 1.

The fitness landscape for a fixed value of variance parameter is depicted in Figure 2. The standard deviation of X is used for computing the variance parameter.

The LCGCDC algorithm involving dynamical fitness function and parameter adaptation mechanism is used for clustering. The convergence of the algorithm is depicted in Figure 3.

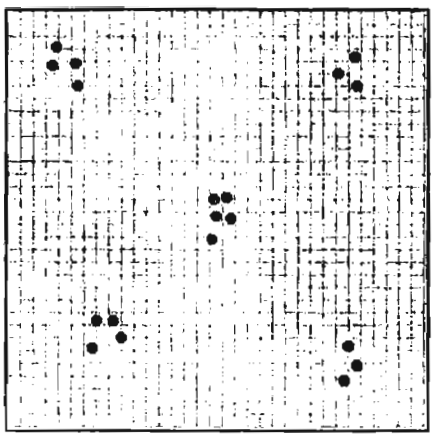

Fig. 1. Data set for clustering and the corresponding Link-Cell division.

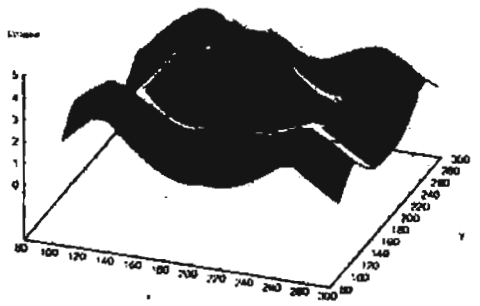

Fig. 2. Fitness landscape using the Gaussian fitness function. Five peaks, corresponding to existing clusters, are detected.

Experiment 2. Link-Cell-Based GCDC vs. Standard GCDC. Twenty different data sets are considered. There are considerable differences between the point distributions, cluster dimensions and distances between clusters among these data sets. LCGCDC is compared with the standard GCDC method.

Using standard GCDC the correct cluster number was obtained for sixteen data sets. For another four data sets a small difference between the algorithm output and the number of real clusters has been observed.

Using LCGCDC without applying final merging and post-processing techniques the correct solution have been obtained for twelve data sets. For another data sets some illegal prototypes have been detected. By applying the final merg-

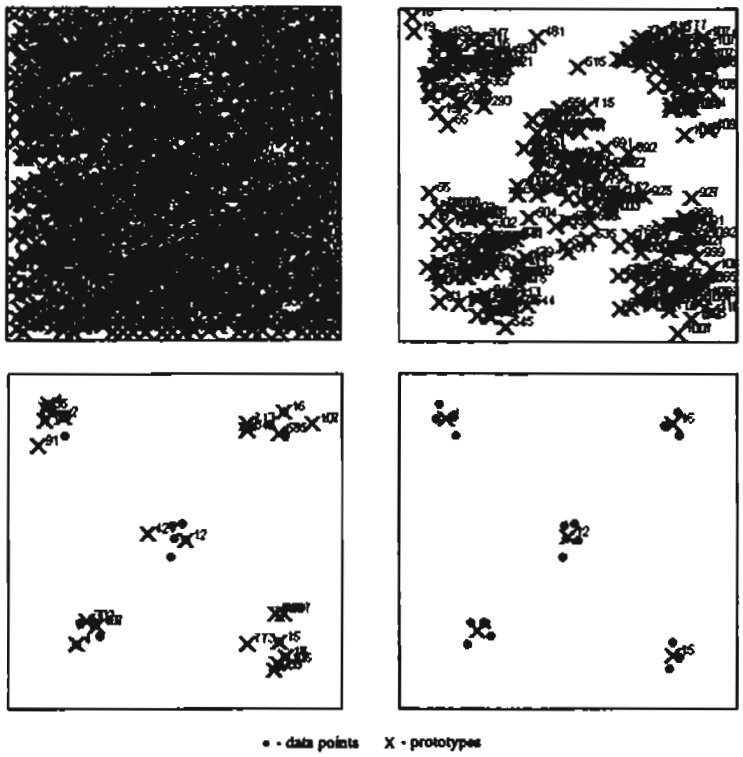

Fig. 3. Convergence of the LCGCDC algorithm. Prototypes obtained after 1, 10, 50 and 150 iterations respectively are depicted.

ing operator four incorrect outputs have been corrected. In two cases three centers have been considered for two closest cluster. The union of these clusters has been interpreted as a separate cluster. In one case a lethal solution remained in final population. By applying the post-processing techniques these small errors have been corrected.

The Link-Cell-based method has been able to determinate the correct cluster number and structure in each situation. There were no essential differences between algorithms in the number of necessary iterations, but there was an essential difference execution time. The LCGCDC method proved to be faster. Using this method the short-range interactions can be easily and efficiently calculated with a smaller computational power.

6 Conclusions

Proposed Link-Cell technique can be successfully used for improving evolutionary multi-modal optimization algorithms. By combining this technique with GC

metaheuristics a new evolutionary model for multi-modal optimization is obtained. This new model is used to derive a new dynamic clustering algorithm.

Link-Cell technique promotes local search and allows new parameter adaptation techniques.

LCGCDC clustering technique can be successfully used for solving clustering problems in a dynamic manner. The algorithm is able to determine the correct number of clusters. The naturally existing cluster structure can be correctly detected by this new algorithm.

With Link-Cell technique short-range interactions can be easily calculated, and the search process becomes faster. Numerical experiments proved that better performances and higher accuracy can be achieved by using Link-Cell-based methods.

References

1. Allen M. P., Tildesley D. J., *Computer simulation of liquids*, Clarendon Press, Oxford, 1987.
2. Deb K., Goldberg D.E., *An investigation of niche and species formation in the genetic function optimization*, Proc. of the 3rd Int. Conf. on Genetic Algorithms, J.D. Schaffer (Ed.), Morgan Kaufmann, San Mateo, CA, (1989), 42-50.
3. Dumitrescu D., Lazzerini B., Jain L. C., Dumitrescu A., *Evolutionary Computation*, CRC Press, Boca Raton, 2000.
4. Dumitrescu D., *Genetic Chromodynamics*, Studia Univ. Babes-Bolyai, Ser. Informatica, **35**, (2000), 39-50.
5. Dumitrescu D., Simon K., *Reducing Complexity of RBF Neural Networks by Dynamic Evolutionary Clustering Techniques*, Proceedings of CAIM, (2003), 83-89.
6. Dumitrescu D., Simon K., *Genetic Chromodynamics for Designing RBF Neural Networks*, Proceedings of SYNASC, (2003), 91-101.
7. Dumitrescu D., Simon K., *Evolutionary Prototype Selection*, Proceedings of ICTAMI, (2003), 183-191.
8. Dumitrescu D., Simon K., *Fitness Functions and Interaction Domain Adaptation Mechanisms for Dynamic Evolutionary Clustering*, submitted to ICCC, (2004).
9. Goldberg D.E., Richardson J., *Genetic algorithms with sharing for multimodal optimization*, Proc. of the 2nd Int. Conf. on Genetic Algorithms, J.J. Grefenstette (Ed.), Lawrence Erlbaum, Hillsdale, NJ, (1987), 41-49.
10. Schreiber T., *A Voronoi Diagram Based Adaptive k-means Type Clustering Algorithm for Multidimensional Weighted Data*, Universitat Kaiserslautern, Technical Report, (1989).
11. Selim S. Z., Ismail M. A., *k-means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality*, IEEE Tran. Pattern Anal. Mach. Intelligence, PAMI-6, **1**, (1986), 81-87.
12. Quentrec B., Brot C., *New methods for searching for neighbours in molecular dynamics computations*, J. Comp. Phys., **13**, (1975), 430-432.