KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 145–152

GENETIC CHROMODYNAMICS. DATA MINING AND TRAINING APPLICATIONS

D. DUMITRESCU[1], KÁROLY SIMON[2], AND ELEONÓRA VÍG[3]

Abstract. To avoid some usual difficulties of standard evolutionary algorithms recently a new multimodal optimization metaheuristics - called Genetic Chromodynamics (GC) - has been proposed. Based on the GC metaheuristics a new dynamic evolutionary clustering technique (GCDC) has been developed. Some applications of GCDC are presented. GCDC is used for gene expression analysis. A GCDC-based method for designing optimal neural network topologies is also presented.

Evolutionary algorithms represent ideal tools for solving difficult optimization problems [3]. Several optimization problems for which classical methods do not work very well or are simply inapplicable can be solved with evolutionary techniques. Evolutionary algorithms can be used for constrained, dynamic, multiobjective and multimodal optimization.

Standard evolutionary algorithms find only one solution, even if the search space is a highly multimodal domain. In order to identify several optimum points special evolutionary models have been proposed. In some cases classical evolutionary multimodal optimization methods, like niching techniques, cannot focus the search on each optimum and find the optimal solutions efficiently.

To avoid some usual difficulties of these standard algorithms recently a new multimodal optimization metaheuristics - called Genetic Chromodynamics (GC) - has been proposed [4]. The model may be used to solve real- world optimization problems including static and dynamic multimodal and multiobjective optimization problems. GC-based techniques can be applied in various scientific, engineering or business fields. Clustering, learning from data, data compression and other data mining problems are very suitable for a GC treatment.

Based on the GC metaheuristics a new clustering technique - called GC-based Dynamic Clustering (GCDC) - has been proposed [7]. Dynamic clustering is a typical multi-modal optimization problem. The problem of cluster optimization is

2000 *Mathematics Subject Classification.* 68Q05, 62H30, 68T05.
Key words and phrases. evolutionary multimodal optimization, dynamic clustering, RBF neural networks, gene expression analysis.

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 [12, 13]), as these typically fix the number of clusters *a priori*. In case of practical problems the number of existing clusters is generally unknown. Dynamic clustering does not require *a priori* specification of the number of clusters. GC-based clustering can be particularly useful to detect the optimal number of clusters in a data set and the corresponding set of useful prototypes [7].

Clustering is a useful exploratory tool in gene expression data, however there are only a few works that deal with the problem of automatically estimating the number of clusters in bioinformatics datasets. GCDC is capable of automatically discovering the optimal number of clusters and its corresponding optimal partition in gene expression datasets.

Solving a problem with a neural network a primordial task is the determination of the network topology. Generally the determination of the neural network topology is a complex problem and cannot be easily solved. When the number of trainable layers and processor units (neurons) is too low, the network is not able to learn the proposed problem. If the number of layers and neurons is too high then the learning process becomes too slow. The main aim is designing optimal topology. In some cases complexity of networks can be reduced by clustering the training data.

In Section 1 the GC metaheuristics and the GC-based dynamic clustering technique are presented. In Section 2 GCDC is used for gene expression analysis. A method for designing optimal RBF neural network topologies using GCDC is presented in Section 3. Some numerical experiments are also described.

1. GC-based dynamic clustering

Genetic Chromodynamics (GC) [4] is a new kind of evolutionary search and optimization metaheuristics. GC is a metaheuristics for maintaining population diversity and for detecting multiple optima. The main idea of the strategy is to force the formation and maintenance of stable sub-populations.

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 the stepping-stone technique each individual in the population has the possibility to contribute to the next generation and thus to the search progress. Corresponding to the local interaction principle the recombination mate of a given individual is selected within a determined mating region. Only short range interactions between solutions are allowed. Local mate selection is done according to the values of the fitness function. An adaptation mechanism can be used to control the interaction range, so as to support sub-population stabilization. Within this adaptation mechanism the interaction radius of each individual could be different.

To enhance GC, micropopulation models 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 using local tournament selection. When the local domain of an individual is empty the individual is mutated.

Within GC sub-populations co-evolve and eventually converge towards several optima. The number of individuals in the 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 optimum, a solution of the problem.

GC allows any data structure suitable for the problem together with any set of meaningful variation/search operators. For instance solutions may be represented as real-component vectors. Moreover the proposed approach is independent of the solution representation.

Based on the GC metaheuristics a new dynamic clustering algorithm - called GCDC - has been developed. This technique is described below.

1.1. **Solution representation.** Corresponding to the proposed GCDC method each cluster is represented by a prototype (cluster center). Each prototype is encoded into a chromosome. The initial population is randomly generated and it contains a large number of individuals.

1.2. **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 subpopulation stabilization an adaptation mechanism is used for controlling interaction domains [9]. For realizing the stepping-stone search principle a micropopulation model is used and it is combined with direct survival competition.

1.3. **Search operators.** 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.

1.4. **Fitness evaluation.** Fitness values of individuals are evaluated using suitable fitness functions. For instance Gaussian functions could be used (see [7, 9]).

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:

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

Parameters of corresponding normal distribution are L_j and γ_j. An adaptation mechanism is used for controlling parameter $\gamma_j, j = 1, ..., m$. In this way a dynamical adaptation of the fitness function is realized.

1.5. **Improving GCDC.** To achieve a better performance final merging and a post-processing methods can be performed [10]. A Link-Cell method can be applied for improving GCDC by promoting local search and deriving new parameter adaptation techniques [11].

2. GCDC for gene expression analysis

Gene expression analysis is of great importance in molecular biology for inferring the functions and structures of a cell since changes in the physiology of an organism are accompanied by changes in the pattern of gene expression.

Gene expression is the process by which a gene's coded information is converted into the structures and functions of a cell. Expressed genes include those that are transcribed into mRNA. The amount of protein that a gene expresses depends on the tissue, the developmental stage of the organism and the metabolic or physiologic stage of the cell. By capturing the cell expression level, biologists can build up a picture of what levels of gene expression may be normal, or abnormal, and what the relative expression levels are between different genes within the same cell [1, 2].

DNA microarray, a recently developed technology, allows thousands of gene expression levels to be measured simultaneously. Data collected by this method is called gene expression data, which after preprocessing (reduction of the noise-level and normalization), forms the data source of our clustering algorithms.

The main characteristics of gene expression data is the very high number of genes (up to 10^6), and the generally small number of samples (< 100). Thus gene expression data is usually represented by a real-valued matrix whose rows correspond to genes and whose columns correspond to conditions, experiments or time points. An element of the matrix represents the expression level of a specific gene under a specific condition.

2.1. **Clustering gene expression data.** Clustering is a fundamental and widely used technique in data analysis and pattern discovery aimed at a better understanding of gene structure, function and regulation. During clustering genes are systematically grouped together according to their similarity in expression patterns. Performing cluster analysis on gene expression data can help detecting gene groups with similar expression patterns, determining the function of new genes,

finding correlation between different groups, understanding gene regulation and cellular processes, observing gene expression differentiation in various diseases or drug treatments, thus digging out biologically meaningful information from genetic data. Multi-gene expression patterns could characterize diseases and lead to new precise diagnostic tools capable of discriminating different kinds of cancers [2].

The desired features of data analysis techniques dealing with gene expression data are robustness, understandability, fastness and automatic detection of the optimal cluster-number. The key challenges regarding gene clustering are the development of methods that can extract order across experiments in typical datasets of size 30000 x 1000, methods which can deal with highly connected, intersecting or even embedded clusters. Boundaries between clusters can be very noisy. There is a need for algorithms that handle effectively these problems.

Several classical clustering and classification algorithms have been applied to gene-expression data from k-means to hierarchical clustering, principal component analysis, factor analysis, independent component analysis, self-organizing maps, decision trees, neural networks, support vector machines, graph-theoretic approaches, and Bayesian networks to name a few. Each method has different advantages depending on the specific task and specific properties of the data set being analyzed. Typically, simpler methods are more robust, while the advanced approaches provide more accurate results [5, 2].

In most clustering methods setting the number of clusters beforehand is necessary; however, the choice of the number K of clusters is a delicate issue, and only a few works deal with the automatic estimation of the number of clusters in bioinformatics datasets.

2.2. **Numerical experiments.** We have chosen a data set for which a biologically meaningful partition into classes is known in the literature. We refer to that partition as the true solution.

RCNS data set [14] contains the expression levels of 112 genes during rat central nervous system development over 9 time points. According to Wen et al. the true partition of this data set contains 6 classes, four of which are composed of functionally related genes. In order to capture the temporal nature of the data, the difference between the values of two consecutive data points is added as an extra data point. Therefore, the final data set consists of a 112 x 17 data matrix. This transformation enhances the similarity between genes.

The GCDC technique provides an optimal clusters-center set containing usually from 5 to 8 solutions; however, the most frequently appearing cluster number is 6.

The slight differences in result sets after multiple runs of the algorithms are due to the stochastic nature of the method. The use of random numbers to pick crossover and mutation locations embed stochastic processes into the algorithm.

Figure 1 shows the average normalized expression pattern over the 9 time points for all the genes in each cluster. These plots are very similar for multiple runs of the algorithm; however the starting dominant ancestor might be different.

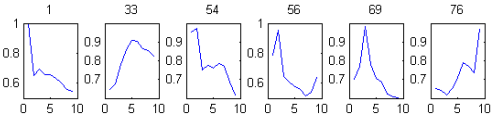

FIGURE 1. Average normalized expression pattern over the 9 time points for all the genes in each cluster. The numbers above each graph represent the indices of the starting dominant ancestors of the different clusters.

The partition retrieved by the algorithm does not correspond entirely to the original partition found by Wen et al., however it also makes sense. It finds a refinement of some initial classes, while others are grouped together. In all cases genes clustered in the same class share similar expression patterns, and a biological interpretation of the classification is also possible.

3. GCDC for designing RBF neural networks

Radial Basis Function (RBF) networks are relatively simple neural networks, especially used for solving interpolation problems [6]. RBF is a feed-forward neural network with an input layer (made up of source nodes: sensory units), a single hidden layer and an output layer. Within RBF networks there is a dependence between the number of training samples and the number of hidden neurons.

Complexity of RBF networks depends on the number of hidden neurons. This complexity can be reduced by clustering the training data. The number of hidden neurons supplies the number of radial basis functions with different centers. It should be favorable the use of training samples as RBF centers, but in some cases this is impossible. If the number of training samples is high, then not all of them might be used (the number of hidden processor units must be reduced). The solution is to consider a single neuron for a group of similar training points. Groups of similar training points can be identified by using clustering methods.

GCDC does not require *a priori* specification of the number of clusters. Therefore GCDC can be used for designing optimal RBF neural network topologies [8].

The number of neurons in the hidden layer of the network is the number of clusters determined by the GCDC method. Cluster centers identified by the GCDC algorithm are used as center parameters for the activation functions. RBF function parameters can be determined according to the cluster diameters. In this way optimal RBF neural network topology can be obtained.

For investigating the performance of the GCDC method a numerical experiment is performed. RBF neural network is used for approximating the function:

$$F_2(x) = 2 \cdot \sin\left(\ln(x) \cdot e^{\cos\left(\frac{x}{2}\right)}\right),$$

where $0 \leq x \leq 9.5$.

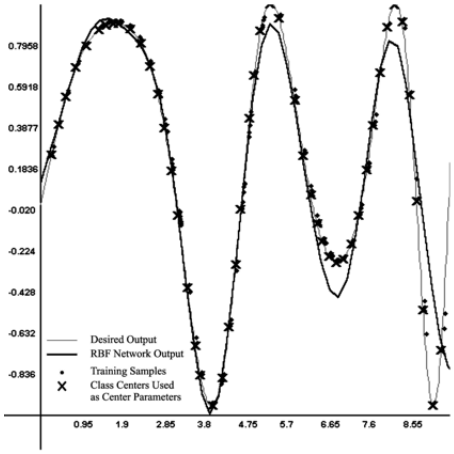

FIGURE 2. 200 training samples organized in 50 clusters, centers determined by the GCDC technique, output of the RBF network after 10000 training epochs.

The network is trained using 10 data sets. Each training set consists of 200 points from the interval [0,9.5]. In each set the points are organized in 50 well-separated clusters. For each set the GCDC method is performed and RBF neural network topologies are created based on the returned results. In 5 cases the number of centers determined by GCDC is 50. In other 5 cases there is a little difference (maximum +4). The *generalization error* is calculated using $M = 400$ inputs (that do not belong to the training set) from the interval [0,9.5]. After training the obtained RBF networks, the mean generalization error is 0.539953496. Satisfactory approximation results are obtained (Figure 2).

The GCDC method has been compared with standard (static) clustering methods. Better results were obtained by using GCDC.

4. Conclusions

Genetic Chromodynamics (GC) a new evolutionary optimization metaheuristics aimed for addressing static and dynamic multimodal and multiobjective optimization problems is described.

A GC-based clustering technique - called GCDC - is proposed. GCDC is used for gene expression analysis and for designing RBF network topology. Numerical experiments indicate the potential of the proposed approach.

References

[1] Attwood T. K., Parry-Smith D. J., Introduction to Bioinformatics, Pretince Hall, 1999Attwood T. K., Parry-Smith D. J., Introduction to Bioinformatics, Pretince Hall, 1999

[2] Baldi P., Hatfield G. W., DNA Microarrays and Gene Expression - From experiments to data analysis and modeling, Cambridge University Press, Cambridge, 2002.

[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), pp. 39-50.

[5] Kohane I. S., Kho A. T., Butte A. J., Microarrays for an integrative genomics, MIT Press, 2003.

[6] Powell M. J. D., *Radial basis functions for multivariable interpolation: A review*, in Algorithms for Approximation, J. C. Mason and M. G. Cox, ed., Clarendon Press, Oxford, (1987), pp. 143-167.

[7] Dumitrescu D., Simon K., *Evolutionary prototype selection*, Proceedings of ICTAMI, (2003), pp. 183-191.

[8] Dumitrescu D., Simon K., Genetic Chromodynamics for designing RBF neural networks, Proceedings of SYNASC, (2003) pp. 91-101.

[9] Dumitrescu D., Simon K., *Fitness functions and interaction domain adaptation mechanisms for dynamic evolutionary clustering*, Proceedings of ICCC, (2004), pp. 132-138.

[10] Dumitrescu D., Simon K., *Post-processing techniques for evolutionary clustering*, Proceedings of the Symposium "Zilele Academice Clujene", (2004), pp. 75-83.

[11] Dumitrescu D., Ferenc Jrai-Szab, Simon K., *Link-Cell method for evolutionary multi-modal optimization. Application in dynamic evolutionary clustering*, Carpathian Journal of Mathematics, 20 (2004), pp. 177-186

[12] Schreiber T., *A Voronoi diagram based adaptive k-means type clustering algorithm for multidimensional weighted data*, Universitat Kaiserslautern, Technical Report, (1989).

[13] 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), pp. 81-87.

[14] Wen X, Fuhrman S, Michaels GS, Carr GS, Smith DB, Barker JL, Somogyi R. Large scale temporal gene expression mapping of central nervous system development. Proc of The National Academy of Science USA., 95 (1998), 334339.

[1] Babes-Bolyai University, Faculty of Mathematics and Computer Science, Computer Science Department

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

[2] Babes-Bolyai University, Faculty of Mathematics and Computer Science, Computer Science Department

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

[3] Babes-Bolyai University, Faculty of Mathematics and Computer Science, Computer Science Department

E-mail address: vig_nora@yahoo.com