The Evolutionary Heuristics Applied to the Analysis of Multidimensional Data

Anna Jasińska-Suwada

University of Technology, Institute of Telecomputing, Cracow, Poland, e-mail: anka@pk.edu.pl

Abstract. This paper describes the pattern recognition system for analysis of multidimensional data, based on natural meta-heuristics. The system consists of tree modules: preprocessing, feature extraction and clustering. Feature extraction module is based on Molecular Dynamic (MD). In clustering are used two natural methods: Simulated Annealing (SA) and Taboo Search (TS). The system is used to analyze an evolving population of individuals equipped with 'genetic codes'. Clustering module extracts groups of data with similar genetic code named clusters and make of possible to observe their geographical localization. The feature extraction verifies the clustering and allows analyzing of clustering patterns, their shapes and the distances between them.

1 Introduction

The pattern recognition system for analysis of multidimensional data, based on natural meta-heuristics consists of a group of procedures, which are used for preprocessing, feature extraction and clustering of multivariate data sets as displayed in Figure 1.

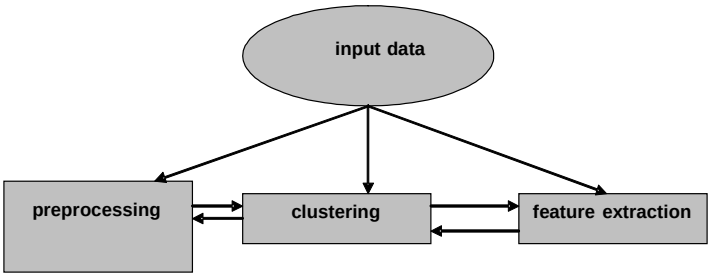


Figure.1. A schema of processing the system for analyzing of the multidimensional data

The multidimensional data are in the form of feature vectors and they represent some objects, which have a specified geographical position. Firstly the vectors are preprocessed both to reduce noisy features and to remove the redundancy of input data. In such a way the most informative features are selected. The role of clustering module is to find the groups of similar elements and

to enable the analysis the clusters of clusters both in feature and in the geographical space. The extraction module transforms data into the feature space. It allows to analyze the effect of clustering, proving the clusters' cohesions, sizes and distances between them. Some of clustering algorithms require giving a number of clusters as input data. The feature extraction procedure enables the verification the number of clusters and to repeat clustering procedure if necessary.

There are a variety of clustering methods based on minimization of certain criterion function (e.g., the widely known k-means clustering [1,2]). These sorts of functions are usually multidimensional and multimodal. Standard minimization procedures enable only the local minimum to be found rather then a global one. So new tools for database clustering are based on evolutionary algorithms, for example using a rule-based genetic algorithm (RBCGA) [3] and multiobjective evolutionary clustering algorithm named MultiObjective Clustering with automatic Kdetermination (MOCK) [4] have been developed.

The feature extraction transforms N-dimensional cluster space into 3-D and then inspects visually the solution obtained from clustering. In the case of an unsatisfactory result, the clustering procedure can be repeated once again or another clustering scheme can be tried.

Each of modules of the system includes programs, which are based on natural methods like genetic algorithms (GA), parallel recombinative simulated annealing (PRSA), molecular dynamic (MD), simulated annealing (SA), taboo search (TS). The preprocessing module is very important because it reduces the complexity of the problem to be solved, but the clustering and feature extraction are also helpful in understanding multidimensional data directly. Therefore in this paper feature extraction and clustering module are discussed.

2 Feature Extraction by Molecular Dynamic

MD (Molecular Dynamic) [5, 6] is useful in optimization multimodal and multidimensional functions and therefore it could be used in feature extraction. The feature extraction known as mapping or feature generation is based on the Sammon criterion [5]. It is another way to reduce the complexity of the problem. It transforms the set of m-dimensional vectors in l-dimensional space. If l=2 or l=3, the output vectors could be easily interpreted by a human and this process is named visualization. It could be reduced to find a minimum of error function, which depends on the distances between elements in both m and 2 or 3-dimensional spaces. The initial configuration of the algorithm is the set of particles scattered randomly in 2 or 3-D space. Each of them represents one point in m-dimensional input space. The particles interact via the two-body potential V_{ij} (1), which corresponds to the error function.

$$V_{ij} = \frac{k}{2l}\left(r_{ij} - d_{ij}\right)^{l} \cdot \left(d_{ij}\right)^{-wl} \tag{1}$$

where: k - stiffness factor
r_{ij} -squared distances between points i i j in $\Re^3$
d_{ij} - squared distances between points i i j in $\Re^m$
w, l - factors; w=0, l=2

Additionally friction dissipates the energy from the system and finally particles stop moving. The particles evolve according to the Newton equation of motion. For slow kinetic energy dissipation

the global minimum of the potential energy is reached and final positions of particles reflect the result of the mapping.

3 Clustering

Clustering [2] is also an optimization problem. The error function J(W,Z) (2) depends on the distances between the points and the centre of the appropriate clusters.

$$J(W,Z) = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}\left\|x_i - z_j\right\|^2 \tag{2}$$

where: n – the number of points
M – the number of clusters
x_i – point i
z_j – the centre of cluster j
W – the matrix of belonging; w_{ij} =1 if point i belongs to cluster j, w_{ij} =0 in other case

To solve this problem two natural methods SA [8] and TS (Tabu Serach) [9] are implemented in the system. In SA each state is represented by matrix W. The changes leading to increasing the error function J are possible with probability p depending on the temperature of the system (3).

$$p = e^{\frac{-\Delta J}{T}} \tag{3}$$

The SA clustering algorithm is show on Table 1. The parameters of the algorithm are: initial temperature T_k, final temperature T_f, cooling factor c, length of Marcov chain.

The Tabu Search algorithm, which minimises the error function J (2) was implemented by Khaled'a Al-Sultan'a [9]. While in this system RTS (Reactive Tabu Search) [10] was implemented. The moving operator is defined as a couple: point and target cluster. The configuration contains matrix describing the points belonging to clusters, the number of iteration and configuration repeating counter. At each step of the algorithm the best move is selected from a set of admissible moves. RTS is memory based search. The old solutions are stored and the escape operator is used to avoid the occurrence of cycles. If configuration repeating counter is larger than a critical value, the space of searching is changed randomly and the memory is cleaned. The tabu list, which includes forbidden moves, is gradually decreased. The RTS algorithm used in clustering is shown on Table 2.

Table 1. The clustering algorithm via SA

```
procedure SA;{
k=0;
initialize matrix W randomly;
count the centers of the clusters Z;
initialize temperature T_k ;      /(high value)
count the error function J_i(W,Z);
repeat
   repeat
      generate new configuration j(W,Z) randomly
      by moving one point to another cluster
      count J_j (W,Z);
      ΔJ_ij=J_j-J_i;
      if  (ΔJ_ij≤ 0)
      then  accept new configuration j;
      else if (random[0,1) < exp (-ΔJ_ij /T_k))
           then accept new configuration j
   until the stability is reached for T_k;
   k=k+1;
   T_k=cT_k-1;
until T_k < T_f ;}
```

Table 2. RTS clustering algorithm

```
procedure RTS;{
generate the initial configuration f^(0); t=0; T^(0);
repeat
   increase the counter of configuration f;
   if counter of configuration >rep
      then {add f  to the set of often repeated;
            chaotic=chaotic+1;
            if chaotic > chaos then {
                  delete the memory
                  make some random moves};
μ = the best move of admissible;
f^(t+1) =μ(f^(t));
modify the set T;
t=t+1;
if J(f^(t))< J_b  then
     {J_b= J(f^(t));
      f_b= f^(t);}
until  stop condition;}
```

4 Tests of the System for Multidimensional Data Analyzing

The described system was used to analyze how the hypothetic population based on Cellular Automata [11] and genetic operators evolves in time, to observe the changes of the genetic codes of individuals.

A motionless population of individuals, each of is equipped in genetic codes, is considered [12, 13]. The individuals are located on a 2-D mesh and the interactions are possible only between the nearest neighbors. The mesh represents a system with limited resources. Features are represented by bit strings of constant length. The maximum life-time of individual is equal to the length of genetic code. During the evolution these codes are read bit by bit at each simulation step. New element of population could be created in free nodes of mesh via genetic operators - crossover and mutation. It is possible only if there are at least two neighbors. 'The crowd' is dangerous for individuals. The element, which has 8 neighbors, is eliminated from the system if the current bit of his genetic code is equal to 0. Otherwise, i.e. for the current bit =1 it survives. The global optimum of in such a define population fills all the nodes of the mesh by elements with genetic codes consisting entirely of '1'.

The population evolves not only in the geographical space (defined here by the 2D mesh) but also in an abstract feature space represented by bits of the genetic chains. The evolution of population is observed both in geographic space - by finding the clusters of similar elements (families) on the mesh - and in the feature space - by transforming multidimensional chains into points localized in the 3D Euclidean space by feature extraction.

At first feature extraction was used to check how the probability of mutation (p_m) influences the diversity of the genetic codes. Fig.4 shows, that for big p_m the point distribution in the feature space is similar to random, represented by a ring in 2-D space. The p_m is less, the individuals are more similar and the number of different kinds of genetic codes decreases. In Figure 2 the colored points represent the groups of identical elements. Their genetic codes are almost filled by '1', they differ only in the two last positions. The green point shows the middle genetic code.

On the rings are located points which represent individuals, for which the number of '0' in genetic codes is constant; their distance from the middle point is the same. The other shapes correspond to families of individuals who have the same values on some position of their genetic code.

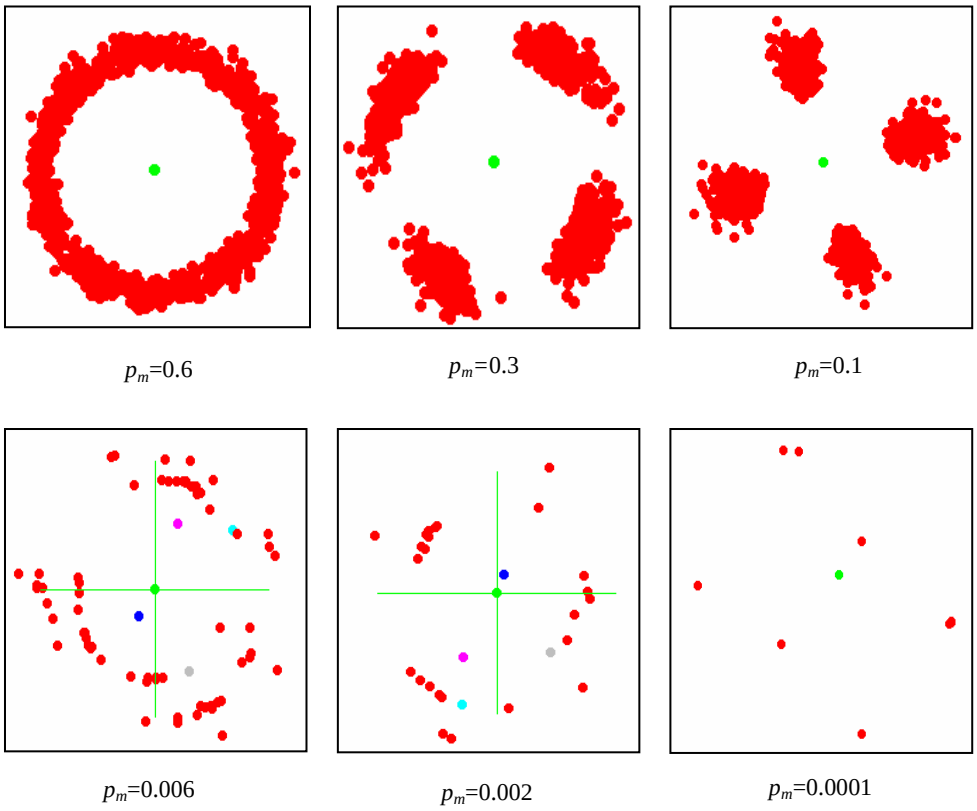
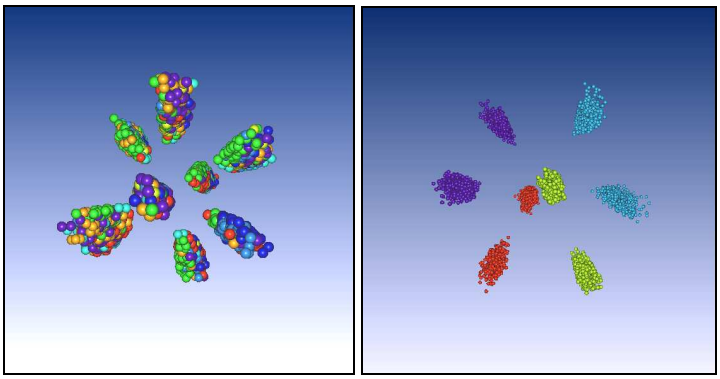
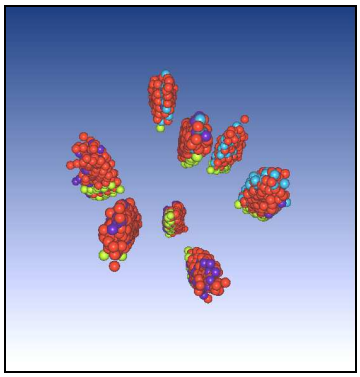


Figure 2. .The effect of clustering and feature extraction in the feature space for some value of p_m

Some modification are introduced to the evolution. One of them is modification of the mutation operator. Its value depends on population density ($p_m=x_n^{30}$) and it is modified in every cycle. The results of clustering performed by means of various schemes yield different results. The analysis of these differences allows for extraction of the most important properties (bits or sub-chains in the 'genetic codes'), which distinguish the population members, 'families' and 'tribes'. After 2000 evolution cycles clustering module has found 4 clusters (Figure 3a). The result from clustering is partly confirmed by the feature extraction scheme (Figure 3a). However, the feature extraction finds twice as many well separated clusters. This means that each cluster extracted by the clustering scheme consists of two sub-clusters, the individuals from the same clusters are not identical. The mutation diversifies the individuals in the clusters. The features, which diversify the population into clusters, are the result of statistical fluctuation and the initial conditions, for different initial conditions different clusters were obtained. None of clustering algorithm found 8 clusters, which clusters would be compatible with the effect of extraction (Figure 3 b, c).



a) 4 clusters b) 8 clusters given by k-means



c) 8 clusters given by SA

Figure 3. The effect of clustering and feature extraction for the modified population

By using clustering schemes and feature extraction procedure we can control the evolution process and detect the most important features responsible for diversification of the population.

5 Conclusions

Evolutionary heuristics could be used to construct pattern recognitions programs, like clustering. The system for multidimensional data analyzing, based on such natural metaheuristisc is suitable for studying some evolution processes. It could be useful for analyzing the changes in the genetic code of population elements. Pattern recognition method can be used for visual analysis of the evolution process in abstract feature space and in geographical space as well.

Bibliography

[1] Ismail, M., A., Kamel, M., S., *Multidimensional data clustering utilizing hybrid search strategies*, Pattern Recognition, Vol. 22, No. 1, pp 75-89, 1989

[2] Theodoris, S., Koutroumbas, K., *Pattern Recognition*, Academic Press, San Diego, London, Boston, 1998

[3] I Sarafis, AMS Zalzala and P Trinder, *A Genetic Rule-Based Data Clustering Toolkit*, IEEE World Congress on Computational Intelligence, CEC02 Proceedings, pp. 1238- 43, IEEE Press, 2002

[4] Handl, J. and Knowles, J. *Multiobjective clustering and cluster validation* Yaochu Jin (editor) *Multiobjective Machine Learning*. Springer Series on Computational Intelligence, 2006

[5] Dzwinel, W., Błasiak, J., *Method of particles in visual clustering of multi-dimensional and large data sets*, Future generation Computer Systems 599, 1-15, 1999

[6] Dzwinel, W., *Informatyczne problemy i perspektywy symulacji metodą cząstek*, Wydawnictwa AGH, Rozprawy monografie, 50, 1996

[7] Siedlecki, W., Siedlecka, K., Sklansky, J., *An overview of mapping techniques for exploratory pattern analysis*, Pattern Recognition, Vol. 31, No. 5, pp. 411-429, 1998

[8] Ingber, L., *Simulated annealing*: Practice versus theory, J. Math. Comput. Model-ling,1993, 18(11), 29

[9] Al-Sultan, K., A., *Tabu Search approach to the clustering problem*, Pattern Recognition, Vol. 28, No. 9

[10] Battiti, R., Tecchiolli, G., *Local Search with Memory: Benchmarking RTS*, Operations Reserch Spectrum 1995

[11] Chopard, B., Droz, M., *Cellular Automata Modeling of Physical Systems*, Cambridge Univ. Press, London, 1998

[12] Jasińska-Suwada, A., Dzwinel, *Evolution of population with limited resources by using genetic operators*, V Krajowa Konferencja Algorytmy Ewolucyjne i Optymalizacja Globalna, Jastrzębia Góra 2001

[13] Jasińska-Suwada, A., Dzwinel, W., *Pattern recognition methods in understanding of evolutionary systems*, Proceedings of the Symposium on Methods of Artificial Intelligence AI-METH 2002, Gliwice 2002