PATTERN MINING AND DISCOVERY ORIENTED TO ARTIFICIAL LIFE①

Liu Jianqin
College of Information Engineering,
Central South University of Technology, Changsha 410083, P. R. China

ABSTRACT The nano-technology requires new methodology to handle difficult problems that involve the information processing, material technology and life phenomena in the nano-world. Concentrating on the synthesis of techniques in scientific frontier fields such as KDD(Knowledge Discovery in Database), evolutionary computation, rough set and logic, a new artificial life model for pattern mining and discovery has been proposed and the corresponding emergent algorithm has been built and implemented. The original contribution of the research work can be summarized in the following two principal respects: (a) pattern mining and discovery for genomic dynamics within the theoretic framework of artificial life; (b) information fusion of multi-paradigm for modeling and building of evolutionary KDD system with rough pattern inference. Through computer experiments the artificial sequence generated by computational processes has matched the evidence convinced by the latest scientific reality. The work is helpful to analyze and build the next generation of bio-nonferrous metal materials in the level of genomics and nano-technology.

Key words artificial life functional genomics pattern mining and discovery

1 INTRODUCTION

Pattern mining and discovery is one of the emerging branches of KDD (Knowledge Discovery in Database) which is a promising field of the frontiers in the international academic community. KDD has been applied in many information processing cases[1-8]. As one of the recent development, the genomics has been paid more and more attentions by the respects of both research and business[9-12]. Recent progresses about the integration of life science and material synthesis can be observed and a significant example is reported and demonstrated in Ref.[13]. The comment in Ref.[14] indicates that modern DNA techniques have successfully been applied to nonferrous metal making such as gold nano-particle processing. A novel artificial life model and corresponding emergent algorithm are proposed in this paper. The key factors such as formal representation through abstract algebraic system, uncertainty mechanism with high dimensional coupled chaotic symbolic dynamics, rough controller schema and object-oriented software architecture paradigm make the modeling be functional and efficient. The multi-disciplinary integration of new subjects such as rough set, evolutionary computation and KDD has laid a strong foundation for the work presented here[15].

2 FORMAL REPRESENTATION AND UNCERTAINTY MODELING

Let the probabilistic measurable space be (S, M, P) and the emergent pattern mining model for uncertainty symbolic processing be $\langle G, E, U, V\rangle$; where S is the whole set of the underlying states, M is the σ-field under the

① Project supported by the Returned Overseas Scholar Foundation of State Education Commission, Open Laboratory Foundation of State Education Commission for Image Information Processing and Intelligent Control, the 21st Oriented Young Scholar Foundation of China National Nonferrous Metals Industry Corporation and Natural Science Foundation of Hunan Province.
Received May 7, 1998; accepted Jun. 8, 1998

meaning of Borel measure, P is the measure for probability, G is the genome set with the meaning of artificial life, E is the emergent computational mechanism for structural dynamics in the level of genomics, U is the nonlinear selection procedure, V is the verification process for pattern discovery.

The genome here takes the form of artificial genome sequence which can be evolved and influenced by the chaotic mechanism embedded in emergence processes. The emergent computational mechanism E consists of two parts (E_f, E_d), where, E_f is the formal description for the functional structure of genome with abstract algebraic language, E_d is the evolutionary dynamics by means of multi-dimensional nonlinear coupled chaotic networks which can be evolved and controlled with the co-evolution mode.

The selection U is implemented by the way of chaos systems to reflect the integration of determined and uncertain information processing. The verification V works as the termination for the circle of machine discovery.

E_f can be defined as a set of abstract algebraic systems which are assigned with the semantic units of field and we get that

$$E_f = (F_0, F_1, \cdots, F_n) \quad (1)$$

where F_i represents the elements of the set and n is the total number of the elements. As the case in this paper we can obtain the following concrete instances.

F_0 is the classical concept of field for the whole domain concerned in the discussion.

F_1 is the subfield that the operations such as addition, multiplication and inverse computation are close for F_0. And F_0 is the extension field of F_1.

F_2 is the algebraic extension field that is satisfied with the following condition:

Let l be a unit in F_2. F_2 is the subfield of F_0 and the extension field of F_1. The element l becomes an upper algebraic element when a nonzero polynominal $L(l) = 0$ and F_2 becomes the algebraic extension field of F_1 when every element in F_2 is upper algebraic element in F_1.

F_3 is a Galois extension field for field F_4 which is the subfield of F_0 and F_4 exists such that it can be a set of fixed points of self-constructed group.

F_5 is an algebraic number field and F_6 is an Abelian extension field such that it can become the separable regular extension field of a field F_7 such that it acts as an Abelian group for the Galois group of F_7. Notice some specific instances for F_6 are assumed.

F_8 is a local field such that a discrete assignment W in F_8 can be defined and F_8 also is complete for W and its remained class field is a limited field.

F_9 is the completion of a field such that its assignment W' makes it become a measurable space and the complete space obtained after the completion procedure for F_9 still be a field in which the complement W' is an extension of W.

The E_f is given as the set of $(F_0, F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9)$ and it is a kind of super-abstract algebraic system. The internal structure of the super-abstract algebraic system is illustrated in Fig.1. The system can constitute an abstract algebraic system which is denoted as (Q, H), where $Q = E_f$ mentioned above and H means the relation among these abstract element.

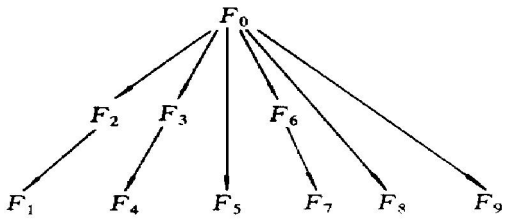

Fig.1 Brief structure of the super-system

E_d focuses on the variation of genome within the evolutionary dynamical genomics and it is expressed as the spatial and temporal chaotic mapping with coupled interaction where emergence may happen among the global hierarchical building based on the integration of competition and cooperation proceeded by the great amount of parallel distributing local cells. The spatial connection of the network is shown in Fig.2. Each chaotic generation unit $K(.)$ is defined as

$$K(n+1) = g_1(K(n), \Psi, \theta) \quad (2)$$

where $n \in I$, Ψ is the parameter set which contains strange attractors, θ is the set of the controllable variables. The initialization for each chaotic generation unit $L(.)$ is implemented by Logistic equation and the temporal evolution processes of the whole network $T(.)$ are selected by the judgment procedure with dynamical thresholding setting produced by nonlinear mapping and chaotic equation

$$S(m+1) = g_2(S(m+1), \Gamma) \quad (3)$$

where $S(.)$ represents the state, $g_2(.)$ takes the combinatorial forms and nonlinear mapping and chaotic equation, Γ is the parameter set. Here the functions $g_1(.)$ and $g_2(.)$ assume as the general form guided by the meta-evolution mechanism specified for the detailed object problem.

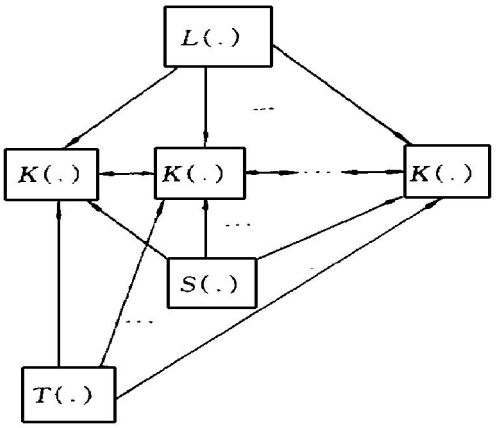

Fig.2 Brief structure of the network

3 SOFTWARE PARADIGM AND CONTROLLER SCHEME OF COMPUTATIONAL PROCEDURE

The structural analysis can provide an efficient and feasible way to pattern mining and discovery system building. As we known, the conceptual description is consistent with the cognitive processes such as the associate memory and semantic inference in thinking and thought especially with respect to the creative activities of wisdom and knowledge. Considering the merits of object description and agent-based programming, a software design paradigm for object-oriented description of pattern mining and discovery system is presented and the flexible arrangement can fit the general framework such as Java virtual machine and network computing environment. The corresponding diagram is given in Fig.3.

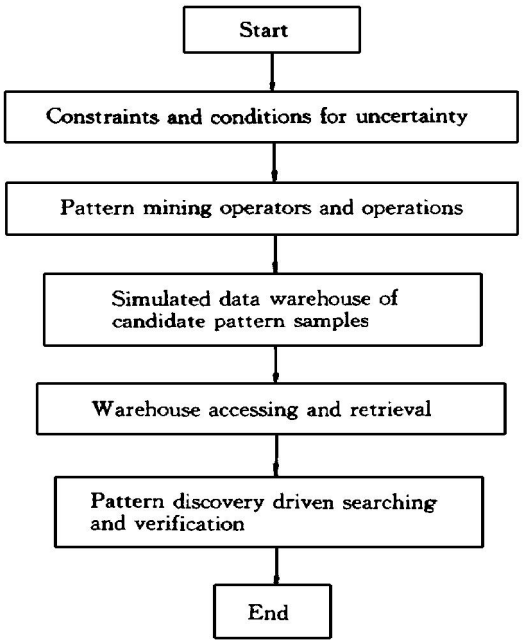

Fig.3 Diagram of pattern mining and discovery

The concerned class set involves the kernel elements and shell components. The kernel set includes the parameterization of abstract algebraic system, symbolic configuration of the evolutionary chaotic dynamics systems and chaotic selection for candidate pattern pool. The shell set consists of operations for warehouse data structure such as creation, maintenance, monitoring, updating and communication (between the different classes for modular architecture), the functional support of the data structure and compiler under certain operating system which is only a selective option for normal developers of the pattern mining algorithm and pattern discovery system. The key classes are listed as follows.

(1) Genome population(type: string);

(2) Structural algebraic description(type: string);

(3) Source data generation(type: real);

(4) Evolution pattern generation(type: string);

(5) Selection action(type: boolean);

(6) Semorphoe labeling(type: integer);

(7) Pattern discovery representation(type: boolean).

For convenience of the transplanting between different working platforms, the numerical calculation parts about chaotic dynamics and symbolic description have been arranged as the compatible forms for SUN Spark II workstation and Pentium PC by standard C source code.

The meta-evolution is performed by the techniques of controlling chaotic mechanism in the underlying evolution procedure. The controlled object is limited to the domain of the chaotic behavioral performance and degree of evolution. The controller structure is displayed in the following schematic diagram (Fig.4).

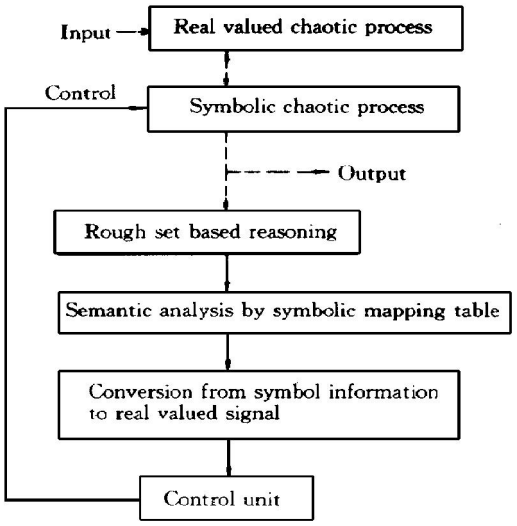

Fig.4 Controller structure

Let the nonlinear mapping from the symbolic space to real space be $Z: B \rightarrow X$, where B and X denote the symbolic and real space, respectively. The rough sets can be defined as the sub-sets of the domain B and the corresponding sub-set for concerned images in X. The upper approximation set and lower approximation set are defined as B_u and B_v for B and X_u and X_v for X. Based on the sets such as B_u, B_v, X_u, X_v, the evolutionary control operators act as the stochastic inference. The principle form of the inference is expressed as the rules:

If (($B_u \subseteq F_u$) and ($B_v \subseteq F_v$)) Then{$f_i = f_d(i)$} (the ith rule)

If (($X_u \subseteq G_u$) and ($X_v \subseteq G_v$) and ($f_i \subseteq K_d(i)$)) Then{$x_i = F_j(x_i)$} (the jth rule)

where F_u, F_v, G_u, G_v, $K_d(i)$ are the decision making area which reflects the domain knowledge. The variable f_i and x_i are the decision outputs and $F_j(x_i)$ is the decision making function.

The rough control strategy is different from the existing uncertainty and impression techniques such as fuzzy set and logic. The difference laid in the place where the membership is avoided and the ambiguity derived from the over subjective hypothesis have been eliminated. The empirical modeling style is beneficial to the unsupervised and non-parametric pattern classification and recognition which the fundamental of pattern mining and discovery is obeyed.

The algorithm has been used to produce the candidate artificial genome for bioinformatics analysis by the computer-aided system in the level of genomics. An experimental result given by the computational procedure with two symbols (U and M) demonstrates the sequence with order of 10 000(cf. Fig.5). And the meaningful result is that a real pattern has been matched with a real sequence reported in Ref.[13]. The pattern is shown as follows,

T-C-G-T-A-C-C-A-G-C-T-A-T-C-C

T-T-T-G-C-T-G-A-G-A-T-C-G-C-G

where this sequence is reported in Fig.2(B) of Ref.[13] at page 1079. Notice that the match between the artificial sequence generated by the computer system and the real sequence has the different meaning from the term "match" mentioned in Ref.[13].

4 CONCLUSION

Up to now, most of the fundamental

Fig .5 The result generated with two-symbol mode

genomics analysis and genome sequencing belong to the domain of the pattern matching methodology. Compared with the increasing speed of the biochemical technical tools, the computer-aided system is imperative and necessary to the upgraded and discovery ability need to be developed with the simulation of creative thinking and modeling for complex systems. The new inverse modeling and approaches presented in this paper have opened an area for the large scale information processing system for tentative, potential and possible "seed" pattern sequences aimed at the discovery of the unknown objects. The artificial result by computer inference fits the "real" sequence. It lays the foundation for automatical mining potential sequences with meaningful biological background and scientific support by machine system.

REFERENCES

1 Fayyad U and Uthurusamy R. Communications of the ACM,1996,39(11) : 24 - 26.

2 Fayyad U and Piatetsky-Shapiro G. Communications of the ACM, 1996, 39(11) : 27 - 34.

3 Glymour C, Madigan D, Preigibon D *et al*. Communications of the ACM, 1996, 39(11) : 35 - 41.

4 Brachman R J, Khabaza T, Kloesgen W *et al*. Communications of the ACM, 1996, 39(11) : 42 - 48.

5 Inmon W H. Communications of the ACM, 1996, 39 (11) : 49 - 50.

6 Fayyad U, Haussier D and Stolorz P. Communication of the ACM, 1996, 39(11) : 51 - 57.

7 Imielinski T and Mannila H. Communications of the ACM, 1996, 39(11) : 58 - 64.

8 Etzioni O. Communications of the ACM, 1996, 39 (11) : 65 - 68.

9 Epp C D. Nature, 1997, 389 : 537.

10 Cohen J. Science, 1997, 275 : 767 - 772.

11 Marshall E. Science, 1997, 275 : 780 - 781.

12 Cohen J. Science, 1997, 276 : 1960 - 1962.

13 Elghanian R, Storhoff J J, Mucic R C *et al*. Science, 1997, 277 : 1078 - 1081.

14 Service R F. Science, 1997, 277 : 1036 - 1037.

15 Liu Jianqin. Artificial Life Theory and Its Applications, (in Chinese). Beijing: Metallurgical Industry Press, 1997.

(Edited by Yuan Saiqian)