An Optimization Model and Approach to the Large Complex System Based on Data Mining

Zhenfeng Jiang*

School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China

Abstract: For the complex system running in multi operation condition, especially the system dynamics changes suddenly from one kind of working condition to another, the controller designed is no longer valid under the premise of the same model. An optimization model and approach are proposed to the large complex system based on the data mining technique. Experimental results suggest that this technology is feasible, correct and valid. Experimental results and trend analysis has many practical significance for improving the performance of large complex systems.

Keywords: Optimization model, optimization approach, data mining, large complex system.

1. INTRODUCTION

The theory and practice have shown, steady, accurate and fast is the basic measure of control system with high quality performance. In order to meet the demand, classical control and modern control theory emerged and developed. In these theoretical frameworks, the design method of control system is based on a single and fixed model, which implied it is based on the operating environment of the system that is either time invariant or slowly time-varying. When the uncertainty affecting system performance is changing in the scope of ‘not too large’, the traditional theory of feedback control, robust control can achieve steady, accurate and fast performance. However, for some complex system, there is greater uncertainty of controlled process and external environment, such as the breakdown of system actuator and sensor mechanism in the process of operation, the unexpected break-in of aircraft from an environment into another, sudden change of external operating environment, and the unknown of part structure and internal parameters. The modeling, measurement and control system of this type of system create challenge to the traditional theory.

Aiming at the problem of dynamic control of large scale and multiple models, Chang et al. put forward a new framework from hierarchical decomposition and coordination control point of view, which is that the large system is decomposed in accordance with the time axis and the parallel algorithm is applied to improve the computing speed, in order to solve the linear quadratic optimal control problem. Giovanini used the parametric linear optimization algorithm to design the optimal control rate and to select the optimal model from the admissible model set. In addition, for overly complex systems, when the number of model is very large, the transient performance of the system will be lowered caused by too much switching.

Lainiotis et al. formed the adaptive control strategy based on the probability weighted method of parameters posteriori, called DUL method. Later on, Li et al. obtained the optimal dual control law in order to solve the LQG problem which had unknown parameters in the state equation. The upside of this method is, although multiple models responding to a plurality of control rate, no switch is needed in the control process and control effect is acceptable, as the following simulation example shows, due to the control applied to the system is designed based on single model which fuses coordinate variables. Decomposition-Coordination control divides the complex system into several simple subsystems by fixing coordinate variables, designs controllers to meet certain performance of each subsystem, and makes performances of all subsystems approach that of the whole complex system by coordinate variables, to achieve the objective of control.

2. PROBLEM FORMULATION

Consider the following discrete-time dynamic systems:

$$z(k)=h\big(x(k)\big)+v(k) \tag{1}$$

$$x(k+1)=f\big[x(k),u(k)\big]+w(k) \tag{2}$$

Among them, x(k) is n-dimensional state vector, u(k) is p-dimensional input vector, z(k) is q-dimensional measurement vector, w(k) and v(k) is the unrelated n-dimensional process noise and q-dimensional process noise respectively in the system operation process, and f and h are n-dimensional and q-dimensional nonlinear functions. The complexity of the system is reflected in the model function f and h which is unknown or known but highly complex, resulting in control and measurement difficulty. Different control will make the system responds differently. And the performance index of the control is the following general function:

$$\Phi = x^T\left(N\right)Cx\left(N\right) + \sum\nolimits_{h=0}^{N-1}\left[x^T\left(h\right)Ax\left(h\right)+u^T\left(h\right)Bu\left(h\right)\right] \tag{3}$$

In which, for a positive semi definite matrix C and a, B is a positive definite matrix, and they all have appropriate dimensions, which can be constant or time-varying, but processing methods are all the same. For the writing simplicity, this paper discusses the steady state, and controls time K which is assumed to change from 0 to N-1, main measuring when system operating condition suddenly change, system model instantly becomes the next model, and this mutation effect on system dynamic is mainly reflected in the transient process. Therefore, in this paper, we assume that K in finite time. For nonlinear stochastic systems (1) and (2) at the K moment, we define real-time information set λ_k as follows:

$$\lambda_k=\left\{u\left(0\right),...,u\left(h-1\right);z\left(1\right),...z\left(h\right)\right\}.$$

The purpose of this paper is to find the control law in the form of $u\left(k\right)=\mu_k\lambda\left(k\right)$, making the performance index (3) the smallest statistically.

$$\begin{aligned}
&\left(\mathrm{P}\right) && \min_{u(k)} E\left(\phi\right)\\
&s.t. && z\left(h\right)=h\left(x\left(k\right)\right)+v\left(h\right)\\
& && h=1,2,\mathrm{L}\ \ ,N\\
& && x\left(h+1\right)=f\left[x\left(h\right),u\left(h\right)\right]+w\left(h\right)\\
& && h=0,1,2,\mathrm{L}\ \ ,N-1
\end{aligned}$$

Among them, E{.} is the expectation operator. It is allowed that the control constraints into $u\left(k\right)=\mu_k\lambda\left(k\right)$ form, because at time k the controller can only know λ_k. Despite that the performance index of problem (P) is a simple quadratic form, it is very difficult to solve the numerical solution of optimal control, because of the nonlinearity of measurement equation and state equation, and changes in working conditions and variety of mutations increase the difficulty to solve the problem. We adopt the following strategy: when system dynamic changes from one operating condition suddenly into another, we will use a set of models to cover controlled object of multiple working conditions. Its mathematical description under multi model framework is as follows:

$$\Omega=\left\{M_i\middle|i=1,2,\mathrm{L}\ \ ,s\right\}$$

Q is a set composed of s models, and the system in the whole control process assumed to have s conditions and each case corresponds to a different model, which reflects the system dynamic mutation. If the ***i^{th}*** model is:

$$\begin{aligned}
&\left(M_i\right) && z\left(h\right)=Hx_i\left(h\right)+v_i\left(h\right)\\
& && x_i\left(h+1\right)=\Phi_i x_i\left(h\right)+G_i u_i\left(h\right)+w_i\left(h\right)\\
& && h=0,1,\mathrm{L}\ \ ,N_i-1
\end{aligned}$$

Among them, $x_i\left(h\right),u_i\left(h\right)$ are respectively the n-dimensional state vector and p-dimensional control vector under the i-th operating condition. Matrix Φ_i, $G_i$with appropriate dimensions, under different conditions, the measuring device is the same. Therefore, matrix H in the measurement equation is fixed. Initial state vector $x_i\left(0\right)$is the white Gaussian noise with mean $\bar{x}_o$ and variance P_0. $w_i\left(k\right)$and $v_i\left(k\right)$ are the unrelated process noise and measurement noise under different operating conditions, and they are unrelated to random initial state $x_i\left(0\right)$ either. $w_i\left(k\right)$ and are the white Gaussian noise with mean 0 and variance W_i and V_i respectively. That is:

$$\begin{aligned}
&w_i\left(k\right):\ \ N\left(0,W_i\right)\\
&v_i\left(k\right):\ \ N\left(0,V_i\right)
\end{aligned}$$

The difficulty of the problem discussed in this paper is, we do not know the system is under which operating condition at the controller design stage, that is to say we do not know which model is the current model among s models. For the LQG problem with unknown parameters, the authors has deeply studied it in the literatures, and obtained meaningful dual control method. The dual controller that designed with this theory can not only achieve the optimal of performance index $E\left\{J\right\}$, but also learn the true value of the unknown parameters. As an attempt, in this paper, the posteriori probability of dual control is used as the coordination variable, and a simple single fusion model is obtained, and then the control law is solved under the LQG framework. This controller can, can make the best compromise between studies in performance and model (operating condition).

3. PROPOSED METHOD

Decision-making tree learning is inductive learning based on examples and also an algorithm approaching discrete function value. It reasons classification rules of expression forms of decision-making tree from a group of disordered and ruleless examples and then uses decisions to analyze new data. In essence, decision-making tree is a process of classifying data through a series of rules. C4.5 algorithm is a type of classical decision tree algorithm. Firstly, ‘split information’ is defined. It can be expressed as

$$split\_info_{\mathrm{A}}\left(D\right)=-\sum_{j=1}^{v}\frac{\left|D_j\right|}{\left|D\right|}\log_2\left(\frac{\left|D_j\right|}{\left|D\right|}\right)$$

Then, gain rate us defined as

$$gain\_oratio\left(A\right)=\frac{gain\left(A\right)}{split\_info\left(A\right)}$$

Hierarchical network is a classical neural network algorithm. All nerve cells in a neural network model are divided into several layers, including input layer, intermediate layer and output layer. Each layer is connected in order. Input in the *I*th layer is only associated with output in (*I*-1)th layer. In neural network, learning process is training process. In other words, in the process of inputting data set in neural network, connection weight among nerve cells is adjusted according to certain method so that network can store data set connotation in the form of connection weight matrix. In this way, when network receives input, proper output may be given.

4. IMPROVED GENETIC ALGORITHM

The improved genetic algorithm is characterized by the extraction and application of heuristic feedback in the whole evolution process. In this paper, the near-optimal solutions obtained throughout the search are analyzed to extract the heuristic feedback, and then the obtained heuristic feedback is used to guide the subsequent search. The computational flow of IGA is shown in Fig. (**1**).

(1) Heuristic Feedback. The first kind of heuristic feedback is called the activity assignment position which is applied to establish a beneficial order for the given activity. A matrix HF_1 with size $N \times N$ is defined for the activity assignment position, $HF_1(i, j)$ denotes the total number of times of assigning the activity i to the j^{th} position among the near-optimal solutions obtained throughout the search. The second kind of heuristic feedback is called the activity assignment person which is applied to establish the beneficial person for one given activity. A matrix HF_2 with size $N \times M$ is defined for the activity assignment person, $HF_2(i, j)$ denotes the total number of times of assigning the activity i to the j^{th} person among the near-optimal solutions obtained throughout the search.

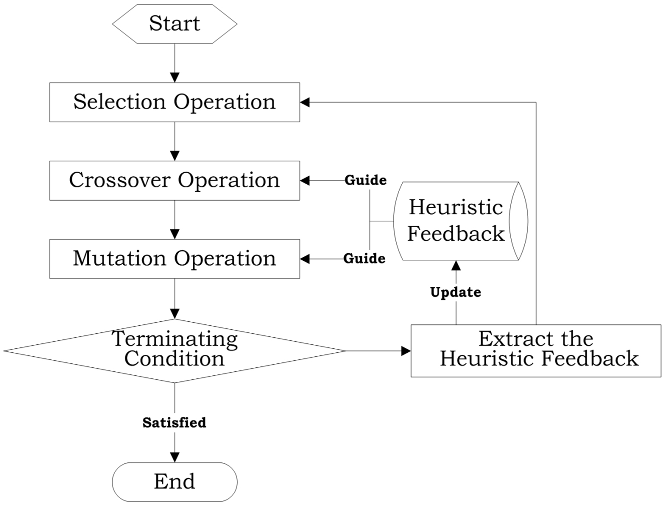

Fig. (1). The computational flow of IGA.

(2) Application of Heuristic Feedback. In IGA, the activity assignment position is applied to guide the crossover operation. The activity assignment position is employed to determine one beneficial position for the given activity. In IGA, the activity assignment person is applied to guide the mutation operation. The activity assignment person is employed to determine one beneficial person for the given activity.

(3) Updating of Heuristic Feedback. After each generation, if the global optimal solution (the best solution from the start) was obtained at this iterative, then the knowledge level will be updated by the following rule, which is based on the optimal solution to accomplish knowledge updating. If the activity i to the j^{th} position among the best solution, then

$$HF_1(i,j) = HF_1(i,j) + 1$$

If the activity i to the j^{th} person among the best solution, then

$$HF_2(i, j) = HF_2(i, j) + 1$$

5. EXPERIMENTAL RESULTS

An organization established a typical complex network information system as the job demand, in which there are 200 assets. Based on the types of information assets and their functions, they can be divided into $M = 10$ kinds. According to statistics, the sum of attacks to the system is $S = 800$, $w_m = 0.1\left(m = 1, 2, \mathrm{L}\ , M\right)$and there are 7 kinds of alternative safety measures. Initial parameters for the decision model can be obtained by expert judgments and data from surveys. According to the financial condition and risk preference of the organization, its size of investment into information safety measures is $\overline{C} = 2700$ million, and through a series of safety measures, its acceptable risk limitation is $\overline{R} = 1000$ million.

Under given constraints, corresponding types of safety measures are taken for different information assets and their implementation costs are adjusted to make the risk and expenses minimum. The proposed optimization model is modeled and solved by genetic algorithm. And relevant initial parameters of genetic algorithm are shown in Table **1**.

Table 1. Initial parameters of feedback genetic algorithm.

Parameter	Meaning	Value
B	Number of subspaces for feasible space in population initialization	3
Q_1	Levels of independent variables in population initialization	5
G	Population size	50
Q_2	Levels of independent variables in crossover operation	5
F	Factors of independent variables in crossover operation	3
σ	Fine-tuned parameter in mutation operation	0.01
SI	Continuous iterations that the global optimal solution has not been improved	50
MI	Maximum iterations	500

The optimal solutions of the decision model can be obtained by model solution. And the best expenses invested into the safety measures are shown in Table **2** (unit: million).

As shown in Table **2**, after costs are invested into safety measures on different information assets, the risk of each asset and the total risk of the whole organization are shown in Table **3**.

It is known from Tables **2** and **3** that, the optimal solutions of the decision model are: the total expense spent in information security risk management is 1713 million, and meanwhile its risk is 913 million quantitatively. It can be seen that, the decisions can meet the requests for costs and

Table 2. Best expenses invested into various safety measures on information assets.

Type of Asset/ Safety Measure	1	2	3	4	5	6	7	Total
1	48	34	0	2	43	3	10	141
2	42	21	0	4	31	25	3	127
3	34	46	42	14	36	30	4	206
4	30	42	7	2	47	18	29	176
5	23	7	21	14	43	43	8	159
6	46	42	10	1	45	13	2	158
7	46	19	18	4	35	8	34	164
8	50	12	6	32	40	38	31	207
9	24	39	32	22	7	22	10	156
10	12	43	26	21	36	31	50	219
Total	356	303	162	117	363	231	179	1713

Table 3. The risk of each asset and the total risk of the whole organization.

Type of Asset	Risk (Million)	Type of Asset	Risk (million)
1	26	6	69
2	55	7	256
3	59	8	121
4	80	9	181
5	55	10	11
Total		913	

affordable risks in organization information safety management, which is taken as a reference for risk decision.

CONCLUSION

An optimization model and approach are proposed to the large complex system based on the data mining technique. Experimental results and trend analysis has many practical significance for improving the performance of large complex systems.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] H. Zhou, W. Cheung, and L. C. Leung, "Minimizing Weighted Tardiness of Job-Shop Scheduling Using a Hybrid Genetic Algorithm", *European Journal of Operational Research*, vol. 194, no. 3, pp. 637-649, 2009.

[2] G. Whittaker, R. Confesor, and S. M. Griffith, "A Hybrid Genetic Algorithm for Multi-objective Problems with Activity Analysis-Based Local Search", *European Journal of Operational Research*, vol. 193, no. 1, pp. 195-203, 2009.

[3] Q. Dai, D. W. Sun, and Z. J. Xiong, "Recent advances in data mining techniques and their applications in hyperspectral image processing for the food industry", *Comprehensive Reviews in Food Science and Food Safety*, vol. 13, no. 5, pp. 891-905, 2014.

[4] G. J. Oyewole and A. E. Oluleye, and E.O. Oyetunji, "Minimizing Maximum Stretch on a Single Machine with Release Dates", *Advances in Industrial Engineering and Management*, vol. 3, no. 1, pp. 1-12, 2014.

[5] J. W. Chung, S. M. Oh, and I. C. Choi, "A Hybrid Genetic Algorithm for Train Sequencing in the Korean Railway", *OMEGA - The International Journal of Management Science*, vol. 37, no. 3, pp. 555-565, 2009.

[6] C. H. Martin, "A Hybrid Genetic Algorithm / Mathematical Programming Approach to the Multi-Family Flow-shop Scheduling

Problem with Lot Streaming", *OMEGA - The International Journal of Management Science*, vol. 37, no. 1, pp. 126-137, 2009.

[7] S. Ross and M. Hann, "Money laundering regulation and risk-based decision making", *Journal of Money Laundering Control*, no. 1, pp. 106-115, 2007.

[8] R. Ronald and S. Yager, "On the Dempster-Shafer framework and new combination rules", *Information Sciences*, no. 2, pp. 93-187, 1987.

[9] S. Agus, N. Sheela, and Y. Ming, "Statistical Methods for Fighting Financial Crimes", *Technometrics*, no. 1, pp. 5-19, 2010.

[10] S. O. Kirnbrough, G. J. Koehler, and M. Lu, "On a Feasible-Infeasible Two-Population (FI-2Pop) Genetic Algorithm for Constrained Optimization: Distance Tracing and no Free Lunch", *European Journal of Operational Research*, vol. 190, no. 2, pp. 310-327, 2008.

[11] M. J. Yao and W. M. Chu, "A Genetic Algorithm for Determining Optimal Replenishment Cycles to Minimize Maximum Warehouse Space Requirements", *OMEGA - The International Journal of Management Science*, vol. 36, no. 4, pp. 619-631, 2008.

[12] Z. L. Liu, "Evaluation on developing level of unban agglomeration derived from resources exploration", *Journal of Applied Sciences*, vol. 13, no. 21, pp. 4702-4707, 2013.

[13] A. B. Garcia, "The use of data mining techniques to discover knowledge from animal and food data: Examples related to the cattle industry", *Trends in Food Science and Technology*, vol. 29, no. 2, pp. 151-157, 2013.

[14] Z. B. Pang, Q. Chen, and W. L. Han, "Value-centric design of the internet-of-things solution for food supply chain: Value creation, sensor portfolio and information fusion", *Information Systems Frontiers*, vol. 17, no. 2, pp. 289-319, 2015.

[15] H. S. Wang, "A Two-Phase Ant Colony Algorithm for Multi-Echelon Defective Supply Chain Network Design", *European Journal of Operational Research*, vol. 192, no. 1, pp. 243-252, 2009.

[16] C. Solnon, "Combining Two Pheromone Structures for Solving the Car Sequencing Problem with Ant Colony Optimization", *European Journal of Operational Research*, vol. 191, no. 3, pp. 1043-1055, 2008.

K. L. Huang and C. J. Liao, "Ant Colony Optimization Combined with Taboo Search for the Job Shop Scheduling Problem", *Computers and Operations Research*, vol. 35, no. 4, pp. 1030-1046, 2008.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014