BASED ON GENETIC ALGORITHM KNOWLEDGE ACQUISITION MODEL

Zetian Fu [1,*] , Le Chen [1] , Yonghong Guo [1], Yongmei Guo [2]

[1] *Department of Mechanical Electrical Engineering,* China JiLiang *University, Hangzhou, Zhejiang Province, P. R. China 310018*

[2] *Department of Computer, Heilongjiang BaoQuanLing Farm Reclamation Industry School, HeGang, HeiLongJiang Province, P. R. China 154211*

[*] *Corresponding author, Address: China Agriculture University Key Laboratory of Modern Precison Agriculture System Integration Peking 100083*

Abstract: In this paper, genetic algorithms and machine learning theory and method of knowledge acquisition for expert system structure is proposed to solve the "bottleneck" problems in the traditional machine learning methods on the basis of AQ, the model of knowledge acquisition based on GA is proposed and its application is used in the Fish Disease Diagnosis reasoning system, the rules is accessed to fish disease diagnosis. Fish Disease Diagnosis and the problems of knowledge accessed into combinatorial optimization are solved.

Keywords: genetic algorithm, knowledge Acquisition, machine learning

1. INTRODUCTION

As expert system application areas continues to expand, the difficulties faced by the combinatorial explosion is increasingly obvious. Construction expert knowledge acquisition system is the "bottleneck" problems, and expertise will have a direct impact on the performance of the whole system. Diagnosis Expert System comes from the knowledge of experts in the field, which is summarized by the expert, it needs repeated exchange between knowledge engineers and experts in the field, and in many cases, often it is very difficult to experience their own knowledge and to make it clear for experts in the field that, especially intuitive problem, the lower the efficiency

of this method. In order to find a suitable for large-scale problems and have self-organizing, adaptive, self-learning ability of the algorithm ,it becomes a research subject goal, people has made a variety of different ways, machine learning is considered one of the most effective means. Traditional machine learning methods are based on pre-established rules and knowledge in the decision taken by the strategy. Therefore it is more difficult for the ever-changing environment of the problem.

Genetic Algorithm as a simulation of natural selection and evolution of the process of a random search algorithm and highly robust in recent years is gradually applied to machine learning system, which has a good performance of the intelligence in determining the coding schemes, fitness function and genetic operator, the algorithm will be used in the evolution of the information, such as self-organizing, adaptive characteristics. At the same time it has also given in accordance with changes in the environment, which can automatically discover the characteristics and environmental laws. Natural selection algorithm eliminates the process of designing one of the greatest obstacles: the need to advance the full description of the characteristics, and describe the different characteristics of the problem algorithm measures to be taken. Thus, we can resolve those structures no one can understand complex issues by using genetic algorithms.

2. BASED ON INDUCTIVE MACHINE LEARNING

The general machine learning framework is shown in Figure 1, the learning system seeks to provide teachers with the concept of a group of samples and background knowledge, identify the description of the concept.

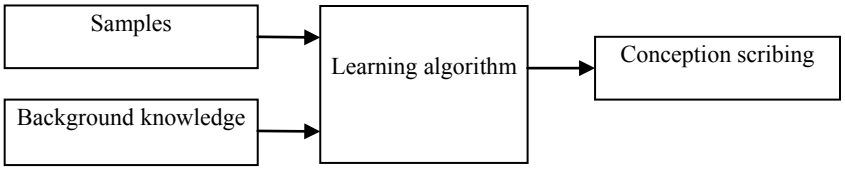

Fig.1 Machine learning framework

Concept can be understood as a group with some of the common nature of the object which is different from other objects. Semantic description of the concept is the basic unit of meaning, which is a significant feature set. Under normal circumstances, in accordance with the concept statement will be divided into the concept and the concept of reasoning. Presented the concept of reasoning is the basis of the analysis of the concept.

To set an example in the database collection, machine learning, known as background sets, $E=D_1\times D_2\times\ldots\ldots\times D_n$ is n-dimensional vector space limited, Dj for a limited collection, its base for mj = | Dj |. E attributes set X = (x_1, x_2,……, x_n), the first j attribute Xj is range D_j. E = <e_1,e_2, ……, e_n> elements in e = <v_1,v_2, ……, v_n> known as examples, $v_j \in D_j$, | E | E for the base said the collection includes examples numbers.

Cases are divided into E will be set of PE and anti-NE:

$PE=\{e^{1+},e^{2+},\ldots,e^{k+}\},e^{i+}=<vi^{1+},vi^{2+},\ldots,vi^{n+}>,K_p\leqslant m$

$NE=\{e^{1-},e^{2-},\ldots,e^{k-}\},ei-=<vi^{1-},vi^{2-},\ldots,vi^{n-}>,K_n\leqslant m$

NE PE and assumptions in n-dimensional attribute set for X, is the same, meet $PE\cup NE=E$, $PE\cap NE=\varphi$,$K_p+K_n=m$，$|PE|=K_p$ ， $|NE|=K_n$.

The concept of learning is a given set of background, the structure contains these statements into assertions, and under the guidance of bias , or the satisfaction of choosing the best summed up assertion that could explain the observed examples of a group concept.

Group related concepts are usually organized into the tree, which can be expressed by the plans or the level of generalization. In the hierarchical structure of a specific level, the concept does not usually intersect, but sometimes they are small differences between large. For example: the concept of a "rotten gill disease" is the "fish diseases," an example of "red skin disease" is a case in point.

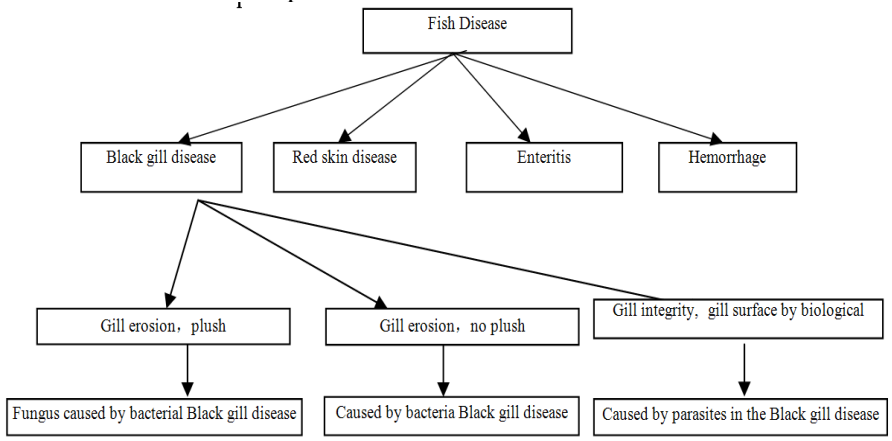

Fig.2 Fish disease concept framework

3. GENETIC ALGORITHMS IN THE APPLICATION OF KNOWLEDGE ACQUISITION

In recent years, many scholars at home and abroad have researched the concept of learning. In these ways, some have used heuristic approach to the

rules to guide the search space, such as ID3 algorithm from the point of studying information theory decision tree; AQ heuristic algorithm using the Star algorithm learning the rules; YAI LS algorithm through an evaluation function, such as learning the rules , High operating efficiency of such methods, but not universal. Some also learn adaptive algorithm which is used to find the optimum rules. GA algorithm is based on Darwin's "natural selection" theory of genetic and biological populations and evolution of the theory of general search methods; it can be used in many areas. In practice, however, GA is not superior to specific areas of the specific algorithm. Therefore, to consider the efficiency and versatility from both sides, people often adopt a more strategic search mechanism, heuristic methods to be combined with the GA, to get a better overall performance of the learning algorithm.

Genetic algorithms for knowledge acquisition may constitute a GA at the core of machine learning system.

3.1 Concept of model

Examples of machine learning is a core areas, in accordance with the knowledge that it can be divided into two categories: decision tree into the decision-making rules and summarized, ID3 as the representative to the former, characterized by training and classification speed quickly, apply to large-scale The learning problems; latter AQ 15 and AE1 represented, characterized by high precision classification, and strong expression of their knowledge, the expert system for automatic acquisition of knowledge and thus in the field of expert systems caused greater concern.

The model, based on examples of the collection E, we have adopted rules that CNF paradigm access to the concept of C, C meet the requirements of the concept of integrity and consistency, and as simple as possible the concept of the length or the shorter the better.

$$if \quad \forall e_i^+ \in PE\ (i = 1,2,.....,\ K_P),\ F(e_i^+, g) = 1 \tag{1}$$

It is said F(g) to meet integrity.

$$if \quad \forall e_i^- \in NE\ (i = 1,2,.....,\ K_N),\ F(e_i^-, g) = 0 \tag{2}$$

It is said F (g) meet the consistency；

To meet the concept of integrity and consistency as binding conditions to simplicity as the goal, the establishment of an integer programming model are as follows:

$$MinZ = \sum_{j=1}^{n}\sum_{l=1}^{m_j} x_{jl} + \omega\sum_{j=1}^{n} x_j$$

$$s.t.\begin{cases} \sum_{j=1}^{n} x_{1l_j j} \geq 1 \\ \sum_{j=1}^{n} x_{il_j j} - \sum_{j=1}^{n} x_{1l_j j} < 0, i = 1,2,..., K_P \\ \sum_{j=1}^{n} x_{il_j j} - \sum_{j=1}^{n} x_{1l_j j} < 0, i = 1,2,..., K_N \\ x_j = x_{j1} \vee x_{j2} \vee ..... \vee x_{jm_j}, j = 1,2,..., n \end{cases} \quad (3)$$

$v_{ij} = d_{jl}, x_{il_j j} = x_{jl}(l = 1,2,..., m)$。 $x_j = x_{j1} \vee x_{j2} \vee ... \vee x_{jm_j}$

if $\forall 1, x_{jl} = 0$ *then* $n_j = 0$

if $\exists 1, x_{j1} = 1$ *then* $x_i = 1$

Where: ω is a big factor punished, the generalω≥2×max{mj,j=1,2,……n}.

3.2 Based on the example of learning the rules GA

The example: The grass carp Hemorrhage and Black gill disease in two cases the rules of study as an example, grass carp Hemorrhage Black gill disease and the symptoms and characteristics of cases Table 1 and shown in Table 2.

Hemorrhage in case of grass carp as are cases of Black gill disease cases as counter-examples to illustrate the concept of learning based on the GA the calculation process and characteristics. In practice, we have set examples of the range of variables and attributes such as shown in Table 3:

Table1 Grass carp Hemorrhage of the sample data Case

ID	Disease	Muscle	Surface	Abdominal	Scales	Head	Fin	Gill	Intestinal
1	Grass carp Hemorrhage	G03	000	000	000	B02	F02	C03	000
2	Grass carp Hemorrhage	G03	000	000	000	B03	F02	C03	000
3	Grass carp Hemorrhage	G03	000	000	000	B07	F02	C03	000
4	Grass carp Hemorrhage	000	A05	000	000	B02	F02	C03	000
5	Grass carp Hemorrhage	000	A05	000	000	B03	F02	C03	000
6	Grass carp Hemorrhage	000	A05	000	000	B07	F02	C03	000
7	Grass carp Hemorrhage	000	A01	000	000	B02	F02	C03	000
8	Grass carp Hemorrhage	000	A01	000	000	B03	F02	C03	000
9	Grass carp Hemorrhage	000	A01	000	000	B07	F02	C03	000

Table2 Black gill disease diagnosed cases of sample data

ID	Disease	Muscle	Surface	Abdominal	Scales	Head	Fin	Gill	Intestinal
1	Black gill disease	000	000	000	000	B01	000	C01	000
2	Black gill disease	000	000	000	000	B01	000	C05	000
3	Black gill disease	000	000	000	000	B01	000	C06	000

Table 3 Examples of existing disease and the collection of attributes Range

No.	x_{ij}	X1	X2	X3	X4	X5	X6	X7	X8
		Muscle	Surface	Abdominal	Scales	Head	Fin	Gill	Intestinal
1	x_{i0}	G03	000	000	000	B02	F02	C03	000
2	x_{i1}	000	A05	000	000	B01	F02	C01	000
3	x_{i2}	G03	000	000	000	B07	000	C03	000
4	x_{i3}	000	000	000	000	B03	000	C05	000
5	x_{i4}	000	A01	000	000	B03	F02	C06	000

Under (3), the issue of model rules for:

$$\begin{aligned}\min\ Z &= x_{10} + x_{12} \\ &+ x_{20} + x_{21} + x_{25} \\ &+ x_{50} + x_{51} + x_{52} + x_{53} + x_{54} \\ &+ x_{60} + x_{61} + x_{64} \\ &+ x_{70} + x_{71} + x_{72} + x_{73} + x_{74} \\ &+ \varpi(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8)\end{aligned} \tag{4}$$

$$s.t. = \begin{cases} x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} \geq 1 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{10} - x_{12} - x_{53} - x_{54} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{10} - x_{12} - x_{52} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{21} - x_{50} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{21} - x_{53} - x_{54} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{21} - x_{52} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{24} - x_{50} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{24} - x_{53} - x_{54} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{24} - x_{52} - x_{60} + x_{61} - x_{64} - x_{70} - x_{72} = 0 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{51} - x_{71} \geq 1 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{51} - x_{73} \geq 1 \\ x_{10} + x_{12} + x_{50} + x_{60} + x_{61} + x_{64} + x_{70} + x_{72} - x_{51} - x_{74} \geq 1 \end{cases} \tag{5}$$

$$\begin{cases} x_1 = x_{10} \vee x_{12} \\ x_2 = x_{21} \vee x_{24} \\ x_5 = x_{50} \vee x_{52} \vee x_{53} \vee x_{54} \\ x_6 = x_{60} \vee x_{61} \vee x_{64} \\ x_7 = x_{70} \vee x_{72} \end{cases} \tag{6}$$

3.3 Based on the binding plan for GA

In general, binding problems can be divided into a binding constraint satisfaction problem (CSP) and is bound by the optimization problem (COP).

For such problems, the issue of the search space does not belong to certain regions include the issue of the search point. Fish diseases from the above model of knowledge acquisition can be seen, it is bound by a non-linear, the problem of how to find such global optimal solution does not exist at present foolproof method.

Optimization of the penalties in the traditional function, and optimize the search process from the point to another point, according to structural constraints of the punishment, punishment will be added to the objective function, so that non-linear programming problems into a series of extreme value to the non-binding question , is the external their early problems that the optimal solution.

To function outside the penalty point method as an example, the model for knowledge acquisition of fish diseases such inequality constrained optimization problem, point penalty function outside the law for the following steps:

Construction penalty function, the original problem into a non-binding Optimization:

$$\phi(X, M^{(k)}) = f(X) + M^{(k)} \sum_{\alpha=1}^{m} \{\max[\ g\alpha(x), 0]\}^{\alpha} \tag{7}$$

Where: the right to punish the second; α punishment for the structural function of the index, its value will affect the function Φ (X, M (k)) in the contour of the binding nature of the general admission α = 2; M For the punishment factor is greater than 0 a progressive series, which should meet

$$0<M(0)<M(1)<\ldots<M(k)<M(k+1)<\ldots)$$
$$\lim_{k\to\infty} M^{(k)} = +\infty \tag{8}$$

Where M(0) for the initial punishment factor, based on experience values (such as from $M^{(0)}=1$).

(2) fitness function is defined as

$$f(X)=C0-\Phi(X,M(k)) \tag{9}$$

where: C0 is a given number, to ensure that f (X) a non-negative, the optimization problem for the sake of fitness into the biggest problem.

Groups of individuals in accordance with the fitness of all sort of mechanism is follows: First, compare the binding of individual fitness Fcom, good adaptation of the top individual, if the fitness of equal value, compared to its optimized fitness Fopt, Good fitness top.

And punishment is usually based on the method, making it possible to the point of total points is better than not feasible. Thus enabling optimization of the process be feasible for the first point, then these potential points better not feasible, and the genetic operation be feasible optimal point. This will enter a viable area and has been optimized, unified, and without changing

the optimum objective fitness to greater than zero. There's no need to set up Fcom and Fopt weight, the use of relatively simple.

Sort after i individuals for the survival chances

$$prob(i)=q(1-q)i-1 \quad (10)$$

One q ∈ (0,1) called the selection pressure, used to control the risk of the individual selected, usually for an average of several times the risk of individual choice.

In order to protect the weak offspring of individuals involved in reproductive capacity and prevent certain anomalies in the evolution of the individual to be disproportionately selected as the father of gene makes the convergence lead to premature convergence, add a counter pNum individual records were selected for the father of the frequency, Article i individuals for the survival chances.

$$prob(i)=q(1-q)i-1(fn)pNum \quad (11)$$

One q ∈ (0,1), can be used to control the individual had been selected as the father of the number of times, pNum bigger, prob (i) the smaller the risk.

（3）Canonical Genetic Algorithms CGA

$\max\{f(b):b\in IB^L\}$

where:$0\leqslant f(b)\leqslant\infty$,$b\in IBl=(0,1)L$,$f(b)\neq const$

Abstract speaking, CGA comes from the following seven components:

CGA=(λ ,L, P0,P',S,C,M)

Where

λ ∈ N for the group in the total number of individuals;

L ∈ N Binary coded for the length of string;

$$P^0=(b_1^0,b_2^0,...,b_\lambda^0)\in I^\lambda, I=IB^L=(0,1)^L \quad (12)$$

P'=（P_c，P_m），where P_c for cross-probability, P_m for the mutation rate;

S：I2→I for the selection；

C：I×IPc，I×I for the cross operator.

In general，Pc and Pm is set their own by the user.

(4) CGA process procedures

Initialization L, λ, Pc, Pm;

Randomly generated initial population, P_{old};

While the termination does not meet the conditions do;

The number of new individual P N = 0;

For j = 1 to λ do fj = f (bj) calculation of the individual fitness;

Optimal solution = max (fj) the corresponding individual;

while N <λ do;

To choose from the operator S P_{old} parents choose;

If PC ≥ Random (0,1) then use a C-operator of two generations;

If Pm ≥ Random (0,1) then use mutation operator M to change the two generations;

The two generations into the new P

N=N+1；

EndWhile；

Pnew=Pold；

EndWhile；

Output optimal values。

（5）Termination criteria

Three kinds of termination criteria can be used: a, using the generation to meet the average value of the older generation and to meet the average ratio of the value of the termination criteria; b, using the frequency of the cycle of the termination criteria; c, the optimal use of individual groups and more and the termination of the same criteria. All of the above criteria can be used alone, but also joint use.

GA solution can be reached through the following conclusions：

IF（Muscle =G03）∧（Head=B03）∧（Fins =F02）∧（Gill =C03）THEN Grass carp hemorrhage

IF（Muscle =G03）∧（Surface=A01）∧（Fins =F02）∧（Gill =C03）THEN Grass carp hemorrhage

Consistent with the actual situation, it can be seen that this method is effective。

4. SUMMARY

Construction expert knowledge acquisition system is the "bottleneck" problems, and expertise will have a direct impact on the performance of the whole system. As Fish Disease Diagnosis knowledge learning is, in fact, a combination of issues, the problem is transformed into knowledge acquisition portfolio optimization problem, fish disease diagnostic knowledge acquisition is used successfully in the model. Genetic Algorithm as a simulation of natural selection and evolution of the process of a random search algorithm, and highly robust in recent years gradually is applied to machine learning system, it has a good performance of the intelligence. A algorithm is proposed in this paper, which is used to solve the concept of access model, it is proved that the method of Fish Disease Diagnosis knowledge acquisition is effective.

ACKNOWLEDGEMENTS

Funding for this research was provided by China Agriculture University Key Laboratory of Modern Precision Agriculture System Integration (P. R. China). The first author is grateful to China Agriculture University for providing her with pursuing a PhD degree.

REFERENCES

D.E. Goldberg .P.Segresi. Finite Markov Chain Analysis of Genetic Algorithm. Genetic Algorithms and Their Applications：Proceedings of the Second international Conference on Genetic Algorithms. 1987，1~8

G .Rudolph. Convergence Analysis of Canonical Genetic Algorithms. IEEE Trans. On Neural Network，1994,5(1)：96~101

H. .Muhlenbein. How Genetic Algorithms Really Work：Mutation and Hillclimbing. in Parallel Problem Solving from Nature，2，Amsterdam , North Holland,1992 ，15~25

J. H. Holland. Adaptation in natural and artificial systems. Ann Arbor：University of Michigan Press. 1975，10

J. R. Koza. Genetic Programming.Cambridge ，MA ：MIT Press, 1992

J. R. Koza. Hierarchical Genetic Algorithms Operation on Populations of Computer Programs. Proceedings of 11th international Joint Conference on Artificial Intelligence，1989

Joe Suzuki. A Further Result on the Markov Chain Model of Genetic Algorithms and Its Application to a Simulated Annealing-Like Strategy. IEEE transactions on System. Man and Cybernetics--Part B：Cybernetics. 1998, 28(1)：95~102T. M. Murdock. et al. Use of a Genetic Algorithm to Analyze Robust Stability Problems, Proceedings of American Control Conference. Boston , 1991, 886~889

Joe Suzuki. A Markov Chain Analysis on Simple Genetic Algorithms. IEEE transactions on System. Man and Cybernetics. 1995，25(4)：655~659

M Srinivas，L.M. Patnaik. Adaptive Probability of Crossover and Mutation in Genetic Algorithms. IEEE Transactions on System. Man and Cybernetics. 1994, 24(4)：656~667

T .Back. The Interaction of Mutation Rate，Selection and Self-Adaption within a Genetic Algorithm. in Parallel Problem Solving from Nature ， 2 ， Amsterdam , Norh Holland,1992 ，84~94

X. Qi，F. Palmieri. Theoretical analysis of evolution algorithms with an infinite population size in continuous space. Part I：basic properties of selection and mutation. IEEE Trans on Neural Networks. 1994，5(1)：102~119