Neurocomputing 174 (2016) 522–529

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A novel attribute reduction algorithm based on rough set and improved artificial fish swarm algorithm

Xin-Yuan Luan, Zhan-Pei Li, Ting-Zhang Liu *

School of Mechatronic Engineering and Automation, Shanghai Key Laboratory of Power Station Automation Technology, Shanghai University, Shanghai 200072, China

ARTICLE INFO

Article history:
Received 7 November 2014
Received in revised form
4 June 2015
Accepted 5 June 2015
Chennai Guest Editor
Available online 1 September 2015

Keywords:
Attribute reduction
Rough set
AFSA
Cauchy distribution

ABSTRACT

Attribute Reduction (AR) is an important preprocessing step for data mining. AR based on rough set is an efficient method. Its reduction performance has been verified to be better or comparable with other methods in large amount of works, but existing reduction algorithms have some problems such as slow convergent speed and probably converging to a local optimum. A novel attribute reduction algorithm based on Artificial Fish Swarm Algorithm (AFSA) and rough set is proposed. For AFSA has a slow convergence rate in the later phase of iterations, normal distribution function, Cauchy distribution function, multi-parent crossover operator, mutation operator and modified minimal generation gap model are adopted to improve AFSA. The attribute reduction algorithm based on improved AFSA and rough set takes full advantages of the improved AFSA and rough set,which are faster, more efficient, simpler, and easier to be implemented. Datasets in the UC Irvine (UCI) Machine Learning Repository are selected to verify the aforementioned new method. The results show that above algorithm can search the attribute reduction set effectively, and it has low time complexity and the excellent global search ability.

1. Introduction

The data mining, also known as knowledge discovery in database, includes extracting/mining knowledge from large amounts of data, discovering new patterns, and predicting the future trends. The rough set theory is an important means of data mining [1–3]. Nowadays, with the explosive growth of web information, the webpage classification faces the great challenge. Wang [4] proposed a classification approach for less popular webpages based on rough set model. Rough set can only deal with attributes of a specific type in the information system by using a specific binary relation [5]. In order to deal with the missing data and incomplete information in real decision problems, Liu [6] presents a matrix based incremental approach in dynamic incomplete information systems. It is very difficult to select significant genes closely related to classification because of the high dimension and small sample size of gene expression data. Rough set has been successfully applied to gene selection, as it selects attributes without redundancy and deals with numerical attributes directly [7]. Rough set has achieved encouraging results which extract target features well and mine data statistical correlation in the financial, industrial production, marketing information system and other fields. Hence, the rough set theory is becoming a hot research spot [8,9].

As known to all, some attributes are useful, while some attributes are redundant or meaningless, which not only occupy extensive computing resources, but also seriously impact the decision making process [10,11]. Attribute Reduction (AR) removes redundant or insignificant information and retains the classification ability of the information system same as before. It is viewed as an important preprocessing step for pattern recognition and data mining. It is also one of the important applications in the rough set theory [12,13]. Rough set method has some advantages: it has explicit stopping criterion and no parameters. And its reduction performance is comparable with other methods or even better [14,15]. Most of researches are focused on attribute reduction by using rough set [16,17].

Some methods of attribute reduction are based on discernibility matrix [1,18,19], some ones on neighborhood rough set model [20–22] and some ones on the variable precision rough set model [23,24]. Minimal reduction problem is even NP (non-deterministic polynomial)-hard problem [18], where the number of attributes is smallest among all possible reductions [14]. Because heuristic algorithms can be used to solve many kinds of NP-hard problems, recently, heuristic attribute reduction algorithm is the main research direction in the field of attribute reduction. They are usually implemented through a certain measure to evaluate the significance of attributes and a heuristic searching strategy [10]. Heuristic attribute reduction algorithms include the algorithm based on attribute significance [25], algorithm based on dependency of attribute [26], and algorithm based

* Corresponding author. Tel.: +86 021 56331563.
E-mail address: liutzh@staff.shu.edu.cn (T.-Z. Liu).

http://dx.doi.org/10.1016/j.neucom.2015.06.090
0925-2312/

on attribute frequency. Specially, the reduction algorithm based on Genetic Algorithm (GA) optimization is proposed by Wrobleswski and other researchers [27,28], and the Particle Swarm Optimization (PSO) algorithm is introduced into reduction algorithm [29,30]. However, such approaches are not difficult to fall into local optimal solution as well. Therefore, searching for a fast and efficient reduction algorithm continues to be one of the major concerns in the field of rough set theory.

According to the above analysis, a novel attribute reduction algorithm based on Artificial Fish Swarm Algorithm (AFSA) and rough set is proposed in this paper. It modifies the heuristic searching strategy and can find the minimal reduction more quickly. AFSA is a new bionic optimization algorithm, which searches for an optimal solution in the target solution space by simulating fish preying, swarming, random and following behaviors [31]. AFSA has some advantages: it can use the target function as algorithm evaluation function directly, and it can get an appropriate solution quickly. Nevertheless AFSA has a slow convergence rate in the later phase of iterations, so several new approaches are adopted to improve AFSA. Normal and Cauchy distribution functions are utilized to optimize visual scope, try number and step. Based on these functions, AFSA can get better balance between convergence speed and solution accuracy. Crossover and mutation operators are introduced to enhance population diversity. Modified Minimal Generation Gap model is employed to retain the elite individual and remove the worst. Preying behavior is removed from swarming and following behaviors to cut down the amount of calculation.

Numerous experiments demonstrate that improved AFSA rough set attribute reduction (IAFSA-RSAR) algorithm has higher search efficiency and lower computational complexity than others.

2. Background

2.1. Basic notions of rough set theory

Redundant attributes add both complexity and overhead of calculation. So, redundant attributes are necessary to be removed, in the meantime, the ability of information classification is kept. Rough set theory is a useful tool of reduction [32].

Let U be a certain set called the *universe*, and X be a certain subset of U. The least composed set in A containing X will be called the *best upper approximation* of X in A, in symbols $\overline{Apr_A}(X)$; the greatest composed set in A contained in X will be called the *best lower approximation* of X in A, in symbols $\underline{Apr_A}(X)$. The *universe* U is divided into three subsets [12,32]

(1) X's positive regions

$$POS(X) = \underline{Apr_A}(X) = \bigcup_{x\in U/X} \underline{A}(x) \tag{1}$$

(2) X's negative regions

$$NEG(X) = POS(\sim X) = \overline{Apr_A}(X) \tag{2}$$

(3) X's boundary regions

$$BND(X) = \overline{Apr_A}(X) - \underline{Apr_A}(X) \tag{3}$$

If $BND(X)$ set is not a null set; the X set is called rough set.

Definition 1. Let U be a certain set called the *universe,* and let R be an equivalence relation on U. The pair $K=(U, R)$ will be called an *approximation space.* Sets $P, Q \subseteq R$, then

$$m = r(P, Q) = |POSP(Q)|/|U| = \sum_{x\in U/Q} \frac{|\underline{P}(x)|}{|U|} \tag{4}$$

$POSp(Q)$ is Q's P positive region in the U space. m $(0 \le m \le 1)$ is called dependency degrees between space Q and P, in symbols $P \Rightarrow_k Q$ where the symbol $|\ |$ means element number of set. If $m=0$, the set Q is completely independent of the set P. If $0 < m < 1$, Q is partly independent of P. If $m=1$, Q is totally dependent on P [33].

2.2. AFSA analysis

Artificial Fishes (AF) are generated by random function in AFSA, and the optimal solution is found out through iterations. At every iteration process, AFs update the maximum fitness value in the bulletin board by preying, swarming and following behaviors.

Preying behavior is AFSA basic behavior, described as formula (5). The X_{inext} value is updated by random function sometimes, this method lets AFSA escape from the local optimal solution. x_i is the AF current state; x_j is a random AF in the visual scope. If $Y_j > Y_i$, the x_i is replaced by x_j directly. Otherwise, the x_i is replaced by another random x_j. Defining the *prey*() function as formula (5).

Swarming behavior is described as formula (6). x_i is AF current state; x_c is the AF in the center of current visual scope; and n_f is the AF numbers in current visual scope. If $Y_c/n_f > \delta Y_i$, it means that there is higher food consistency in the center of current visual scope, meanwhile, it is not too crowded then AF goes toward a step to the center. Fitness value $Y(x_i)$ is calculated. $Y(x_i)$ is compared with the value in the bulletin board, it replaces the latter if it is greater than latter.

Following behavior is described as formula (7) x_i is AF current state; x_{max} is the AF with the greatest food consistence among companion AFs in the current visual scope. n_f is the AF number in current visual scope. If $Y_{max}/n_f > \delta Y_i$, this means that x_{max} has the higher food consistence in the center of current visual scope, meanwhile, it is not too crowded, then AF goes toward a step to x_{max}. The fitness value $Y(x_i)$ is calculated. $Y(x_i)$ is compared with the value in the bulletin board; it replaces latter if it is greater than latter.

Swarming and following behaviors help AFSA locate the optimal solution. For the AFSA, if visual scope is greater, global search ability will be stronger and convergence speed will be faster. Oppositely, if visual scope is smaller, the local search ability will be stronger. If the updated step is greater, convergence speed will be faster. If the updated step is smaller, convergence speed will be slower, but the precision will be higher [31].

$$X_{inext} = \begin{cases} X_i + \text{Random(Step)}\dfrac{X_j - X_i}{\|X_j - X_i\|} & \text{if } (Y_i < Y_j) \\ X_i + \text{Random(Step)} & \text{else} \end{cases} \tag{5}$$

$$X_{inext} = \begin{cases} X_i + \text{Random(Step)}\dfrac{X_c - X_i}{\|X_c - X_i\|} & \text{if}\left(\dfrac{Y_c}{nf} > \delta Y_i\right) \\ prey() & \text{else} \end{cases} \tag{6}$$

$$X_{inext} = \begin{cases} X_i + \text{Random(Step)}\dfrac{X_{max} - X_i}{\|X_{max} - X_i\|} & \text{if}\left(\dfrac{Y_{\max}}{nf} > \delta Y_i\right) \\ prey() & \text{else} \end{cases} \tag{7}$$