

Application of modified artificial bee colony algorithm in health condition diagnosis

YuKe Zhang

School of Zhejiang Sci-Tech University, Zhejiang 310018, China

358793325@qq.com

Keywords: Health condition, Feature selection, Artificial bee colony algorithm, Bayesian.

Abstract: In the medical field, it is very important for doctors to make effective and correct decision-making. In order to improve the accuracy of doctors' diagnosis and avoid the misdiagnosis of doctors' intuition, subconscious and incomplete knowledge. ABC-NB algorithm is used in the field of chronic disease diagnosis to improve the diagnostic efficiency and reduce the chance of misjudgment. The artificial bee colony algorithm based on improved scale factor is applied to the selection of chronic disease characteristics, and the data are dimensioned, the redundant and irrelevant features are removed, the convergence speed is improved, and the algorithm is applied to search the global optimal solution. Then, the eigenvalues of the pre-processed data are trained and learned to generate the Bayesian classifier to construct the prediction model. The prediction module displays the diagnostic results for medical staff to assist in the diagnosis and decision making. Experiments show that the model has good flexibility and robustness, can have a stable calculation of the probability of diagnosis of chronic diseases, and it is effective for the diagnosis of medical staff.

1. Introduction

Now, the rapid development of information at the same time, people are increasingly concerned about their own health, more and more people want to be more understanding of their physical condition. In recent years, with the tide of the Internet of Things continue to advance, more and more health-related Internet equipment into the public life, a large number of health data stored in the Internet at the same time, health data showed explosive growth, the system database stored Of the relevant health information data appears to be more important. But because of the data in a decentralized, discontinuous form, only by manual reading is not enough to analyze the massive data from the more useful information. If you can use modern computer technology to filter a variety of data, analysis, and extract the professional medical-related clinical experience and disease prevention diagnosis knowledge, construct an auxiliary measurement model, to provide services and assistance, you can meet many people Own health concerns.

Since a person's health data is judged by a number of indicators, there are many unrelated characteristic attributes, and it is a primary task to choose from a large number of indicator data and to find valuable information for health diagnosis. Feature selection refers to the problem of selecting the relevant characteristics that produce the most measurement results. The feature selection can eliminate the irrelevant or redundant features and retain the effective features so as to achieve the optimal subset from the original data set. In many practical problems, there are no universal methods for searching and filtering feature variables. The common algorithm heuristic search sequence forward selection algorithm has the disadvantage that only the eigenvalues are increased and the eigenvalues are not reduced. There are similar random search strategies Of the feature selection algorithm, the disadvantage is that when there are more features, the algorithm in the implementation process will consume a lot of time, practicality is not high. In this paper, an artificial bee colony optimization algorithm is proposed for feature search and selection. Experiments show that the

overall measurement rate of the model is stable at 80% or more, effectively improving the diagnostic efficiency.

2. Bee colony algorithm for feature selection

2.1 Algorithm steps

Assuming that the spatial dimension of the solution is that the number of bees and the observed bees and the nectar are equal, the number is given. The standard ABC algorithm regards the process of solving the optimization as a search in the D-dimensional search space. The position of each nectar represents a potential solution to the problem, and the amount of nectar of nectar is corresponding to the fitness of the solution, and a leading bee corresponds to a nectar. x_d^{\max} and x_d^{\min} correspond to the upper and lower boundaries of the dimension search space.

$$x_{id} = x_d^{\min} + rand(0,1)(x_d^{\max} - x_d^{\min})$$

In the search phase, lead the bee in the corresponding first honey around the formula according to the following formula to find a new source:

$$w_{id} = x_{id} + \phi(x_{id} - x_{kd})$$

And $i = 1, 2, \dots, D$, d is a random integer between $[1, D]$, ϕ is a random number evenly distributed over interval $[-1, 1]$, $[k \neq i]$. In each of the leading bee to complete the operation, fly back to the honeycomb will be the quality of the nectar, distance and other information to share the observed bee, each observation bee according to the probability of choosing a honey source, the probability formula:

$$p_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j}$$

If the observation of the bee based on the information to share the honey to select the honey, the same will also produce a new honey around the source, and according to the choice of nostalgia.

In the search process, fit_i is the possible value of the solution X_i . When all the leading bees and watchers have searched for complete space after t iteration, if the fitness value of the nectar is not raised when the threshold is reached, the nectar is discarded and the corresponding bee Switch the role into the investigation bee, the investigation of bees in the search space randomly generated a new source instead of X_i , m for the number of iterations, the process is as follows:

$$X_i^{m+1} = x_d^{\min} + rand(0,1)(x_d^{\max} - x_d^{\min}), m \geq limit$$

$$X_i^{m+1} = X_i^m, m < limit$$

In order to maintain universality and without loss of generality, the fitness of the solution is evaluated in the bee colony algorithm according to the following formula.

$$fit_i = 1/(1 + f_i), f_i \geq 0$$

$$fit_i = 1 + abs(f_i), f_i < 0$$

The bee colony algorithm is compared with other group intelligence algorithms, and the role switch is the unique mechanism according to the information of the honey source. The bees are in the reconnaissance bee, followed by the bee and lead the bee three kinds of different roles to switch between, through the sharing of information to find the best match. In the process of ABC algorithm optimization, the role of different bees is different: lead bee used to maintain good solution; follow the bee used to improve the convergence rate; reconnaissance bee used to increase the ability to jump out of the local optimal solution. The process of searching for the optimal nectar is the process of finding the global optimal solution of the objective function.

3. Feature Classification Model Based on Naive Bayesian Network

3.1 Naive Bayesian classifier

Naive Bayesian classification is a relatively simple classification algorithm and a very low error rate, Bayesian refers to the Bayesian theorem, simply refers to the characteristics of independent conditions. Naive Bayesian's thought is based on the following: Solving the probability of the given class to be classified in each category, the category with the highest probability is considered to be the category to which the item to be classified belongs.

For a problem of solving the classification, according to the Bayes theorem, the probability formula of the class to be classified x belongs to class y is:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

In this formula, x is a feature vector, and the x dimension is M . Since the characteristic condition is independent in the naive Bayesian classification algorithm, the above formula can be expressed according to the full probability formula:

$$p(y=c_k|x) = \frac{\prod_{i=1}^m p(x^i|y=c_k)p(y=c_k)}{\sum_k p(y=c_k)\prod_{i=1}^m p(x^i|y=c_k)}$$

Where $y_i \in \{c_1, c_2 \dots c_k\}$ belongs to a class of K , as long as it is estimated that the sample data feature x^i corresponds to the conditional probability on each class. The probability of category y_i in all categories can be calculated from the training set, and the data characteristics correspond to the criteria for each category. Independent probability vectors can also be obtained from the training set.

3.2 Parameter Estimation

The generation of the classifier is the main work of this stage. The stage is mechanical. By inputting the characteristic attribute and training sample, the proportion of each category in the training sample and the corresponding proportion of each characteristic attribute under different categories are calculated and Generates and outputs the classifier based on the recorded results. The program can be calculated automatically according to the formula discussed above.

Build a training set:

$$TrainingSet = \{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\}$$

3.3 Classification

According to the training model obtained by the training, we can calculate the non-classified instance X through the probability formula $p(y=c_k|X)$, and then calculate the molecular part, the specific calculation steps are as follows:

1. Calculate the probability that the instance belongs to class $y=c_k$:

$$p(y=c_k|X) = p(y=c_k)\prod_{i=1}^m p(X^{(i)}=x^{(i)}|y=c_k)$$

2. Determine the category y to which the instance belongs:

$$y = \arg \max p(y=c_k|X)$$

3. Health diagnosis model and implementation process

First of all, to prepare the naive Bayesian classification, according to the specific situation to determine the characteristics of the property, the input is to be classified sample data, output and the quality of the classifier related to the characteristics and training samples. Randomly initialize a honey source, each honey source in the population represents a subset of the feature, the feature subset of the feature selection.

For each iteration to determine whether the algorithm satisfies the optimal solution, if it is satisfied, then the optimal solution is output; otherwise it will continue to iterate, here we introduce the scale adjustment factor to find the nectar to make it more close to the honey process.

In this paper, we set the maximum number of iterations of the global maximum iteration and the local maximum iteration of the threshold to limit the algorithm to ensure the termination of the algorithm. After the iteration, the optimal solution of the honey source, the source of the representative of the characteristics of the subset is the optimal feature subset. The weight of the population, age, body weight, body mass index, body age, body fat rate, bone mass, muscle content, moisture content, visceral fat grade, basal metabolism, protein content.

It is shown that the GA algorithm is more likely to produce premature convergence as the percentage of data sets increases, and the ABC algorithm continually updates the scale adjustment factor in the iterative process, making the process of searching the nectar in space closer to the colony. The behavior of natural honey can be used to guide the colony to converge to the optimal solution in the better direction, which reduces the dimension of the feature subset, thus optimizing the characteristic attribute input in the classification training stage and improving the accuracy of the model classification.

4. Conclusions

With the rapid development of the Internet today, the medical field produces massive clinical data, although the doctor can intuitively view and analyze each clinical data for diagnosis, but on the one hand because of the huge amount of data, beyond the time and energy can accept Range, can only see part of it; on the other hand the doctor based on clinical experience and expertise may cause the diagnosis is not accurate [10-11]. Based on the analysis of the principle of the colony algorithm and the naive Bayesian classifier, this paper proposes a chronic disease diagnosis model based on the improved scale factor and the Naive Bayesian classifier, which has good flexibility and robustness , But also improve the classification performance of the model, reducing the training time. Doctors can be based on the results of the classification of the model, and its own clinical experience and expertise combined, so not only save time, but also can improve the accuracy of diagnosis. In a subsequent study, the model will be applied to different disease areas while continuing to improve diagnostic accuracy.

References

- [1] Suguna N, Thanushkodi K. A Novel Rough Set Reduct Algorithm for Medical Domain Based on Bee Colony Optimization[J]. Computer Science, 2010.
- [2] Uzer M S, Yilmaz N, Inan O. Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification.[J]. Scientific World Journal, 2013, 2013(11):419187.
- [3] Karaboga D, Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm[M]. Kluwer Academic Publishers, 2007.
- [4] Rish I. An empirical study of the naive Bayes classifier[J]. Journal of Universal Computer Science, 2001, 1(2):127.