Advances in Engineering Research, volume 113
International Forum on Mechanical, Control and Automation (IFMCA 2016)

A PSO-based Generic Classifier Design and Weka Implementation Study

Hui HU[1, a] Xiaodong MAO[1, b] Qin XI[1, c]
[1]School of Economics and Management, East China Jiaotong University, P.R.China
[a]hh24895@163.com, [b]maoxd@ecjtu.jx.cn [c]xiqin@ecjtu.jx.cn

Keywords: Classifier; PSO algorithm; Weka platform; instance-based learning.
Abstract. Aiming at dataset with different feature attributes and multi nominal decision categories, a new design algorithm using PSO for instance-based learning classifier is put forward which can facilitate to make a better classification decision. On the basis of classes and interfaces in Weka platform, this improved algorithm is implemented and added in the platform. Taking two datasets as for test cases, not only is the classifier compared with optimizing different objective functions, but also is evaluated with other classifiers. Comparisons demonstrate that the proposed algorithm can get more effective classifying results and maybe considered a generic classifier.

Introduction

As an important analysis approach in Data Mining, classifying is to build a classifier or classification model which can be used to evaluate model and category prediction for test cases or new data instances[1]. There're a great many classifying algorithms [2], such as Decision Tree, Bayes Network, Supported Vector Machine, Neural Network, Lay Learning and Statistical Learning etc. As for these algorithms, advantages and disadvantages exist simultaneously in them [3,4]. For instance, some classifying algorithm can gain a high computation speed at the expense of data type restriction and sensitive to noise data while other algorithm may get a more accurate prediction by longer modeling and terrible explanation. Among them, instance-based learning is considered to be a lazy learning algorithm, whose fundamental is to minimize the distance between identical category instances and maximize the distance between different category instances by taking an instance as a point in N-dimension space and a feature attribute as a dimension[5].

Particle Swarm Optimization (abbreviated to PSO) is put forward to search global best value by tracking current optimum by J.Kennedy and R.Eberhart in 1995 as a kind of swarm evolutionary algorithm[6]. Owing to its easy implementation, high accuracy and fast convergence, PSO is receiving more and more attention in academia and industry and widely applied in various fields. So is for data mining, like feature selection, data partition, classifying, optimization[7] etc.

In this paper, in order to solve a classifying problem only for binary decision attribute, an improved PSO-based classifying algorithm is proposed to design a generic classifier for multiple attribute decision on the basis of different adaptive objective functions firstly. Then, the implementation of this algorithm in Weka platform is introduced and the classification results are compared and demonstrate the algorithm's effectiveness by testing two typical datasets.

PSO-BASED CLASSIFYING ALGORITHM

PSO Algorithm

The principle of PSO algorithm can be depicted as followed. A solution of the optimal problem is considered to a particle in multi-dimension space and each particle has a corresponding adaptive value which is decided by objection function; flying at a certain speed in the searching space, a particle can adjust its velocity and position by tracking its current local best value and the populations' global optimal value so that the optimal solution of the problem may be found.

Classifier Design based on PSO

Particle code representation is a prerequisite for PSO-based algorithm. In this study, an instance in dataset is considered to be a particle while feature attributes constitute of multi-dimension space except for decision attributes. For continuous feature attribute, the distance between particles is

defined the adaptive value of a particle, which can be computed by Euclidean distance, Minkowsky distance etc. While for discrete feature attribute, the distance can be transformed into simple matching coefficient, Jaccard coefficient etc.

On the basis of basic PSO algorithm described above, an improved PSO-based classifier's algorithm is proposed as below:

Input: PSO algorithm basic parameters include maximum iteration number, maximum inertia weighted coefficient, minimum inertia weighted coefficient and learning factors; dataset to be classified, the number of datasets with different categories (assumed to be equal) etc.

Output: Classified results of the dataset.

The detailed procedures for this algorithm is as followed:

Step1: According to different decision attribute, datasets with different categories are got from the given dataset.

Step2: Produce initial velocity matrix of particles.

Step3: Dataset with different categories are sampled from the given datasets with different categories.

Step4: A population dataset is composed of datasets with different categories according to the given number of datasets.

Step5: For different category dataset, compute each sample's adaptive value and the decision attribute's value.

Step5.1: Compute the distance between each instance in the dataset with different category and the distance between each instance in the population dataset.

Step5.2: Compare the distance between each instance in the dataset with different category and population dataset. If it is minimal for some category and identical with the given category, classifying is correct and recorded.

Step5.3: Save each instance's accuracy and can be got through classifying correct number divided by total sample number.

Step6: Set initial samples as the local best value for different category datasets. Select the maximum accuracy datasets with different categories as the global best values of the different categories.

Step7: Compute the flying velocity of each instance with different category.

Step8: Compute the position of each instance with different category and keep the category unchanged.

Step9: Compute the local best position of each instance.

Step10: Using equation (4) to compute the global best position of each instance.

Step11: Check the termination condition. If not ended, return to Step7. Otherwise, go to next step12.

Step12: Compute the global best value and product output.

IMPLEMENTATION AND TEST ANALYSIS

PSO-based Classifying Algorithm Implementation in Weka

As an open-source platform for data mining, Weka covers a wide range of algorithms and tools in data preprocessing and data modeling and can fulfill various data mining tasks which include attribute selection, regression, classifying, clustering and associate rule etc. One advantage of Weka is that it's very convenient to improve existed algorithms or add new algorithms once following Weka's programming rules such as inheriting classes and interfaces.

On the basis of the steps described above, this improved PSO-based classifying algorithm is implemented in Weka using Eclipse6.5 development software. This new classifying class which is set in weka.classifiers.functions inherits from abstract Classifier and implements OptionHandler, Observer interfaces and its main methods are designed as below:

(1) PSOClassfier( ): defaulted constructed function

(2) BuildClassfier(Instances data): for classifying

(3) void InitDataset(Instances data): initialize data from the given datasets

(4) double Calculation_Fitness(Instances data,Instances swarm): compute sampe data adaption value

(5) double Calculate_GBest(Instances data): get the global best value

(6) double Calculate_LBest(Instances data): get the local best value

(7) double Calculate_Velocity_Nest(Instance data,GBest): get the particle's velocity according to global best value)

(8) double Calculate_Position(Instance data,velocity): get the particle's position according to it's velocity.

This new improved PSO-based classifier is implemented as Fig.1.

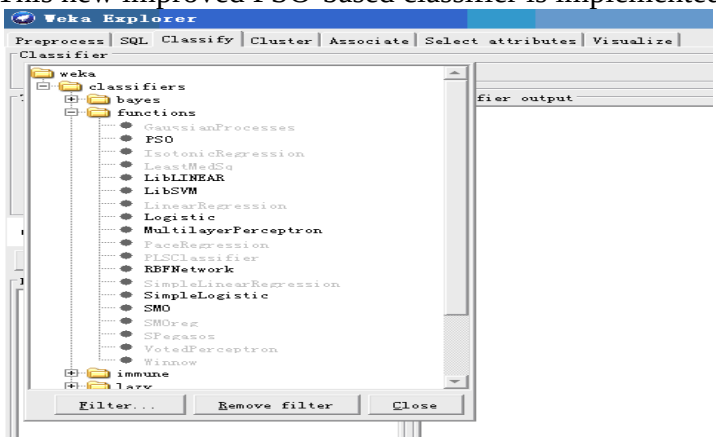

Fig. 1 A PSO-based classifier implantation result in Weka

Case Test Analysis using different objective functions

In order to analyze the algorithm parameters' influences on classifying effectiveness, classifying results are compared among different objective functions. The case test datasets are originated from reference[10] which is concerned with breast cancer diagnosis. There are 9 feature attributes and breast cancer class is decision attribute which has two values (one is malign and the other is benign). We choose two hundred data cases as the test data. The number of benign tumor is 100 and so is for malign tumor. Other data is similar to the reference[7]. The fitness functions are set Euclidean distance, cosine distance, Minkowsky distance(cube root). The classification results are got using the PSO-based algorithm as Table 1 shows. It can be seen that different objective functions have distinct influences on the classification results which demonstrate that the best is Minkowsky distance function, the worst is cosine distance function and Euclidean distance function falls between these two functions.

Table 1 Classification results base on different distance function

Objective function	Breast cancer	benign	Malign
Euclidean distance	benign	100	0
	malign	5	95
Cosine distance	benign	98	2
	malign	5	95
Minkowsky distance	benign	100	0
	malign	4	96

Case Test for Multiple Decision Attributes

The above test cases' decision attribute has only two values. To further confirm this improved PSO-based classifier's performance, we select UCI dataset Iris to make a study. There're 3 values for decision attribute: setosa, versicolor and virginica. The parameters in our algorithm are set as followed. Maximum iteration number is set to 50; maximum weighted inertia coefficient is 0.9; minimum weighted inertia coefficient is 0.1; learning factors c1, c2 are set to 2.

The classification results are compared with other algorithms such as J48 in Weka and BayesNet in Table 2.

Table 2 Classification results base on different classifier algorithms

Classifier algorithm		setosa	versicolor	virginica
J48	setosa	49	1	0
	Versicolor	0	47	3
	Virginica	0	2	48
BayesNet	Setosa	50	0	0
	Versicolor	0	44	6
	Virginica	0	5	45
PSO-based	Setosa	50	0	0
	Malign	0	48	2
	virginica	0	3	47

As can be seen from Table 2, although different classifiers have different classification results, PSO-based classifier is the best one.

CONCLUSIONS AND DISCUSSIONS

Instance-based learning is one of important classifier design approaches in data mining. As a widely used evolutionary algorithm, PSO is receiving more and more attention in data mining processes. Aimed at classifier design for instance-based learning, an improved PSO-based algorithm is proposed to optimize multiple objective fitness functions in classification decision problem with multiple attribute values. Furthermore, implementation of this PSO-based generic classifier is briefly described in Weka by introducing main methods and screenshot result. Then comparison analysis is carried out using two test cases and results demonstrate that this PSO-based algorithm has a better performance compared to existing classifiers.

What's more, it's necessary to point out that this algorithm is based on basic PSO in our paper. Though a great many improved PSO are put forward such as hybrid PSO, prey-escape PSO etc., basic PSO algorithm has the almost same classification results as those other improved PSO algorithms as long as suitable parameters are selected.

Acknowledgements

This work is supported PhD Start-up Fund of East China Jiaotong University 26441001.

References

[1] Luan Li-Hua, Ji Gen-lin. Study on Decision Tree Classification Technology, Computer Engineering, 2004,30(9): 94-96.
[2] Wang Miao, Chai Rui-min. An Improved feature selection method for Decision Tree, Computer Engineering and Application, 2010,46(8):127-129.
[3] Wang Hui-qing, Chen Jun-Jie and Hou Xiao-jing. Study on feature selection method on Decision Tree Classification, Journal of TaiYuan Technology College,2011,42(8):27-129.
[4] David W., Dennis Kibler, Markc and K. Albert. Instance-Based Learning Algorithms, Maching Learning, 1995,6(1):37-66.
[5] Kennedy J, Eberhart R Eberhart. Particle Swarm Optimization, Proceeding of IEEE International Conference on Neural Networks, Perth, Western Australia, 1995:1942-1947.
[6] A. Erdeljan, D. Capko, S. Vukmirovic, D. Bojanic, V. Congradac. Distributed PSO Algorithm for Data Model Partitioning in Power Distribution Systems, Journal of Applied Research and Technology, 2014,12(5):947-957.

[7] Alejandro Cernantes, Ines marla Galvan, Pedro Isasi. AMPSO : A New Particle Swarm Method for Nearest Neighborhood Classification, IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics. 2009,39(5):1082-1091.