J. Basic. Appl. Sci. Res., 3(1)885-890, 2013

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Predicting Missing Attribute Values Using Cooperative Particle Swarm Optimization

Mohammad Hossein Norouzi Beirami, Mohammad Hossein Nejad Ghavifekr, Rahim Pasha Khajei

Department of Computer Engineering, Osku Branch, Islamic Azad University, Osku, Iran

ABSTRACT

In recent years, the problem of predicting missing attributes values has noticed in data mining and discovering knowledge from source of data. Simplest procedure encounter with attribute values is ignoring them which in this case we will miss valuable information.
Different methods have been proposed with this problem. Most of this methods use decision rules for predicting missing attribute values. In this paper, a new method which using Cooperative Particle Swarm Optimization offer for predicting attribute values. This method utilizes data records for convergence to missing attribute values without extracting data relations. We do not need knowledge of professional person for detection relation between data in this method. Proposed algorithm has been done on whether casting of Tabriz which has gain for 50 years that shows accuracy 98.45 percent of missing data.
KEYWORDS: Data Mining, Prediction Algorithm, Replacing Values, Fitness Function.

1. INTRODUCTION

One of the problem of data analysis and extraction of knowledge from mass of information is the problem of missing data [1]. There are various reason for missing attribute value such as deleting information, fields not being exactly predicted in set of old data, data elimination as a result of transferring from sensors , etc [2]. Several methods have been reported for solving the problem of missing values including using common value of the missed attribute, using the concept of the common value of the missed attribute, assigning any possible value that represents a valid concept, ignoring the missing attribute, considering the missed value as a special value, etc [1, 2, and 3]. In this paper, we will present a new method to face this problem. This method uses Cooperative Particle Swarm Optimization for prediction missing values. Proposed algorithm is a new method in this context, that utilizes existence records in dataset to predict missing value and there is no need to knowledge of expert human.

In the rest of the paper, different parts are organized as follows. In second section, we will present existing methods for prediction of missing values. In third section, we will introduce Particle Swarm Optimization and sort of them. In section four, the proposed algorithm is introduced for prediction of missing values. In section five, we explain assessments that are done and finally, in sixth section, we introduce the result of experiments and the future works.

2. PREDICTION AGORITHMS FOR MISSING VALUES

Rough Set Theory

For the first time, Rough Set Theory introduced by powlock in 1982 and since then we witness the theoretical and applicable development of this topic in the world. This theory has a new mathematical view to the concept of “rough” and “ambiguous”.

The base of this theory is the knowledge and information that is in the posses of each observation of the studied set that is called reference set. In these methods, each rough concept is determined by a pair of exact concepts that are called lower approximation and upper approximation of the rough concept.

The lower approximation is composed of all observations which definitely belong to the rough concept and also, upper approximation consists of all observations which probably belong to the rough concept [1]. The difference between lower and upper approximations is border area of that concept. Creating lower and upper approximation in rough sets establishes the principle of information analysis in next steps. Two categories of rules are extracted for prediction. Certain rules are achieved from lower approximation. In fact, lower approximation is definitive and certain state for prediction. Possible rules are also achieved from upper approximation and upper approximation has a probable state in decision makings [2, 3].

*Corresponding Author: Mohammad Hossein Norouzi Beirami, Department of Computer Engineering, Osku Branch, Islamic Azad University, Osku, Iran. Email: Noroozi@iauosku.ac.ir

Frequent Items Set

In this method we use association rule mining technique. In data mining context, this technique is exploited to discover relations between items in a mass dataset. In the production of association rules, frequency of an item appearance in a dataset according to relation between items is depicted with a parameter called "Support". Using "support" of different items, we are going to compute the possibility of appearance of one or some items in the current transaction. Hence, frequency of items can be used for predicting of missing attribute values [2]. In order to produce frequent items set in association rules algorithm, having a "support" more than a particular threshold, the item is added to an item set in order to generating the association rules. This process continues until all items of the dataset are investigated [4, 5].

Decision rules generation

Classic methods for extracting decision rules are a two stages process. In the first stage, we find a solution for covering all existence samples. Set of rules are achieved either directly via inference of conjunctive rules or indirectly through extracting from decision tree [6].

It is usually inferred single rule each time in direct solution and the case that is covered by that rule is eliminated. This process is repeated. In the second step, the coverage rule set is converted to a smaller structure and it is better to select independent rules set using statistical tests. The rules achieved from this method can be utilized to predict missing attribute values [7, 8]. Furthermore, classification models can be exploited to predict missing attribute values.

3- PARTCLE SWARM OPTIMAIZATION

Standard Particle Swarm Optimization

Particle Swarm Optimization is a part of parallel searching population based algorithm which starts to work with a group of random answers, then in order to achieve to optimized answer in problem space, it continues searching by updating location of particle. Each particle is identified in a multidimensional manner by two vectors namely X_{id} and V_{id} indicating location and velocity of d_{th} dimension of ith particle respectively. In each stage of population movement, location of each particle updates with two best of values. The first value is the best experience has achieved from particle up to now and it is shown as g_best. In each repetition, algorithm updates speed and location of new particle based on equation one and two after finding high two values [9].

$$v_{id}(t+1) = w.v_{id}(t) + c_1.rand_1(p\_best_{id}(t) - x_{id}(t)) + c_2.rand_2(g\_best_{id}(t) - x_{id}(t)) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

In equation (1), W is Inertia coefficient which decreases linear and it is in the range of [0-1].

C_1 and C_2 are the acceleration coefficient that are selected in the range of [0-2] and most of the time are considered 1.49 [9, 10, 11, 12]. $Rand_1$ and $Rand_2$ are random numbers in [0-1]. Also, for preventing from Divergence of algorithm final value of each particle speed is limited to one interval. $v_{id} \in [-v_{\max}, v_{\max}]$. The algorithm continues until either a certain value of convergence is reached or the algorithm runs for a particular number of repetitions. Equation (2) updates location vector of new particle considering its new speed and its present location.

Cooperative Particle Swarm Optimization

Cooperative particle swarm optimization is an extension to standard PSO in which several swarms work together to find optimized answer. Different methods are proposed for this task [23]. In [13, 14] several swarms are used for search. One of the swarms is the main swarm and other swarms are subsidiary swarms supporting the main swarm in order to reach to best answer. In [15, 16] instead of using a single swarm to solve an "n" dimensional problem, "n" swarms are used to solve n number of single dimensional problems [22].

Extended Cooperative Particle Swarm Optimization

In [12] we proposed a new extension for PSO algorithm that first of all decreases the problems of falling in to local optimum trap. Secondly, increases the speed of converging, moreover it eliminates randomness of initial particle initialization. On the other hand, the main goal of cooperative algorithms that is the parallel computing capability is achieved. Figure 1 indicates the pseudo code of CPSO algorithm.

The proposed algorithm includes two main parts. The first part consists of n independent swarms of particles. In the first stage, searching space is divided to n parts and each swarm is responsible for searching a single part. After this division, PSO algorithm is applied on each part. In first stage, 20 percent of the desired iterations are executed on each part. Then, first stage is finished and g_best of each part is considered as primary solutions of the main swarm. Now, the main swarm is initialized by n particles while their initial values are g_best of the first stage for each part. In other words, g_best of each swarm in the first stage is the local optimum for its independent area. Main swarm is initialized using local optimums computed in the previous stage. Now that initial solutions have been set PSO is applied on the whole search space for remaining 80 percent of iterations.

CPSO Algorithm

Define g_best: array [1..n] k-dimension vector // for all swarms
Divide search space to n independent areas
Initialize n k-dimensions PSOs: $P_j \quad j \in [1..n]$ **// n is the number of swarms and each swarm works in one area**
For all swarms
  For i=1: t // t is 1/5of all iterations in simulation
    Apply PSO // each swarm is limited in own area
    Update g_best array for each swarm
  End
End
Initialize k-dimensions PSO with g_best values // swarm have n particles
Select the best area with minimum fitness for next search
For i=t+1: end of simulation
  Apply PSO for search space or selected area
End

Figure 1: Pseudo code of CPSO

4- THE PROPOSED ALGORITHM FOR PREDICTING MISSIED VALUES USING CPSO

Considering the section 2, most of existence methods for predicting missing values are methods based on rules. Some rules for existence set of data are discovered. Then missing values are predicted using these rules. High volume of accurate data is needed in order to achieve rules with high reliability.

In this section, we present a different method for predicting of missing values which is based on PSO algorithm. Considering that we introduced CPSO in section 3, we utilize this algorithm in missing values prediction. Figure 2 shows missing values prediction algorithm by using CPSO.

1. Firstly, records with missing values should be extracted from dataset to achieve a dataset with no missing value record.
2. k-means algorithm is applied on the new dataset which have no missing value to cluster dataset into *C* clusters [17, 18, 19, and 20].
3. One of the records having a missed value is selected to be used in prediction.
4. Minimum and maximum value of the missed attribute is searched in the dataset.
5. Array [1.. h] which values are in the range of [minimum, maximum] of the missed value.
6. Fitness function with h members is created. This function is a table with h number of the record with missed value in which Array [1.. h] are copied in missed field and other fields are the same as of the record fields as shown in Table1.
7. C number of swarms is generated and each swarm is initialized using one of the clusters.
8. For a particular number of iterations (20% of the whole iterations) PSO is applied individually on each cluster by individual swarms.
9. Now, there are C numbers of g_bests achieved from clusters. We should generate a new swarm with C particles which are initialized by the found g_bests.
10. Again PSO should be applied (for 80% of the whole iterations) on the search space.
11. The final result of the above step is going to be instead of missing value.

Figure 2: pseudo code of CPSO algorithm for predicting of missing value

As far as we know, this algorithm is applied by using PSO in this context for the first time and important stage of the task is selecting fitness function. Table 1, shows how to choose this function on aerologic data. In fact if missing field is average temperature rate then its range of variation should be [-20.5, 30.1]. We copy rest of the fields (non missed value fields) of the record exactly from the record with its original values. Values of the missed fields are initialized with values that start from the minimum value found from the dataset to the maximum value. This range is divided to h (number of records, for example 20 in Table1), and each field is added with this value. The purpose of our algorithm is to being converged to one of the records of the fitness function (here Table1 is the fitness function). The more the number of records are the more the accuracy of the task is. On the other hand, the speed of task is decreased.

Table 1: Fitness function for aerologic dataset

Raw	Average of Temperature (Celsius)	Maximum Temperature (Celsius)	Minimum Temperature (Celsius)	Average of Pressure	Precipitation (mm)	...
1	-20.5	26	15	50	0.2	...
2	-17.6789	26	15	50	0.2	...
3	-14.8579	26	15	50	0.2	...
.	.					
.	.					
.						
.	.					
.	.					
.	.					
.	.					
19	30.27895	26	15	50	0.2	...
20	33.1	26	15	50	0.2	...

5-EVALUATION OF THE PROPOSED METHOD

Considering aforesaid, we can say common points of the existing methods is that most of them are rule based. It has to be extracted some rules for each dataset. Moreover, as another point missing values should have similar values to be replaced with missing values. As a contrary to the existing methods, in the proposed algorithm, none of these points is necessary. In fact, it is tried to converge some random records to record with missing value. In our method the achieved answer that is the result of convergence doesn't essentially exist in the dataset, however the answer is the result of convergence to the missing value.

Considering the fact that the existing methods include a vast area of applications and success ratio of these algorithms depends on prediction accuracy of missing values. So, optimized method is the method that can predict data with higher accuracy. For example, in [7], for associative rule with support of 10%, the prediction accuracy of 96% is achieved.10 percent support for rules has less comprehensive for existence data. In [6], different methods based on decision rules and association rules have been compared. These methods have been applied on dataset of several cars. Final answer is 73% of prediction accuracy. In [8], a decision rule based algorithm has been experimented on standard data of UCI and has predicted about 65 percent accuracy data.

Results of Prediction Using The Proposed Algorithm

To experiment proposed method we used data which has been collected for 50 years from East Azerbaijan aerology organization, Tabriz airport station and in all we received 17434 reliable records. Fields related to this data and variation range of each field are shown in table 2.

Table 2: dataset of aerology

Raw	Fields	Variation range
1	Average Temperature	[-20.5,33.1]
2	Minimum Temperature	[-25,33.5]
3	Maximum Temperature	[-14,42]
4	Precipitation	[0,63]
5	Average Pressure	[843,879]
6	Minimum pressure	[840,878]
7	Maximum pressure	[846,888]
8	Average humidity	[9.8,98.2]
9	Minimum humidity	[1,96]
10	Maximum humidity	[16,100]

Considering that the number field of this table (dataset) is 10, hence each particle has 10 dimensions. Also, we divide dataset to 6 separate clusters. We consider the number of records of fitness function, 20. We repeat the algorithm 1000 times that 200 iterations are for initial sets and 800 iterations are for final step that is the whole search space with founded g_bests of each cluster.

To do the experiment we randomly select a record from the dataset and delete the value of one of the fields of that record. This experiment is applied on temperature fields. Now, we apply algorithm for predicting the missing value. Achieved result is compared with the original value. It is noticeable that the eliminated field doesn't

contribute in the algorithm or any calculation. Table 3 is the results of 1000 times of independent simulation runs which shows average difference of predicted and original values.

Considering the results of table 3, we can say proposed method has a successful performance in predicting missing values. In this evaluation, the missed value has been selected from one of the minimum, maximum or average temperature fields. These data were in the range of [-25 42.3] centigrade according to the dataset. In this evaluation we tried to find 1000 missed values in 1000 simulation run. In 35.89% of predictions there is only a difference of .5 between results and original data and for 17.62% of prediction we have difference between 0.5 and 1 unit. 31.73% of predictions have difference between 1 to 2 units and finally, for 14.76% of predictions we have more than 2 units difference. Totally we can predict missing values with accuracy of 98.45% and there is an average difference of ± 1.04 unit for all predicted values. By comparison with results of existing methods which were introduced in first section, these results have predicted missing values with high accuracy and therefore the proposed method shows outstanding performance.

Table 3: Predictions results using CPSO

Difference Range	Predictions in each range	Average of difference	Accuracy percentage
[0 0.5]	35.89%	± 0.25	99.62%
[0.5 1]	17.62%	± 0.77	98.86%
[1 2]	31.73%	± 1.40	97.92%
> 2	14.76%	± 2.51	96.27%
All data	100%	± 1.04	98.45%

6- CONCLUSION

At the end, we can say most of the existence methods are based on rules and we have to extract rules from datasets. Quality of these rules and learning them has a direct effect on the quality of missing value prediction. But proposed algorithm tries to make set of random data to converge into the record with missing value. Furthermore, we have to emphasize that the proposed method doesn't need an expert knowledge in order to predict missing values.

The single problem of the proposed method is execution time of the algorithm. Since this method is based on population and should be repeated for several generations to get to results so, it is time consuming in comparison with methods based on rules.

We can say that the proposed method is a new research area in this context. The important phase of the algorithm is the selection of fitness function in order to evaluate the results.

Considering the result of simulations it is demonstrable that the proposed method has gained better results and having an accuracy of 98.45 percent in predicting missing values on real data is a promising proof.

REFERENCES

[1] Jerzy W. Grzymala-Busse "Rough Set Strategies to Data with Missing Attribute Values" Proceedings of the Workshop on Foundations and New Directions in Data Mining associated with the third IEEE International Conference on Data Mining, November 19–22, 2003, Melbourne, FL, USA, 56–63.

[2] Jiye Li, Nick Cercone, "Predicting Missing Attribute Values based on Frequent Itemset and RSFit" ACM, 2006

[3] W. Jerzy and G. Busse and A. Y. Wang, "Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values", the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences, 1997.

[4] W. Jerzy and G. Busse and M. Hu. "A comparison of several approaches to missing attribute values in data mining", In Rough Sets and Current Trends in Computing Springer-Verlag Berlin Heidelberg pp. 378−385, 2001.

[5] M. Kryszkiewicz, "Association rules in incomplete databases", Springer, 1999.

[6] M. Sholom and L. Nitin, “Decision rule solution for data mining with missing values”, USA, 2000.

[7] L. Jiye and C. Nick, “Comparisons on different approaches to assign missing attribute values”, 2006.

[8] C. Blake and E. Keogh and C. Merz, “Uci repository of machine learning database”, University of California Irvine, 1999.

[9] J.Kennedy and R. Eberhart, “Particle swarm optimization”, Proceedings of IEEE International Conference on Neural Networks, vol. 4, 1995.

[10] Sheybani, M. and Meybodi, M. R., "PSO-LA: A New Model for Optimization", Proceedings of 12th Annual CSI Computer Conference of Iran, Shahid Beheshti University, Tehran, Iran, pp. 1162-1169, Feb. 20-22, 2007

[11] Sheybani, M. and Meybodi, M. R., "CLA-PSO: A New Model for Optimization", Proceedings of 15th Conference on Electrical Engineering (15th ICEE), Volume on Computer, Telecommunication Research Center, Tehran, Iran, May 15-17, 2007

[12] Norouzi Beirami, M. H. and Meybodi, M. R., "Cooperative Fuzzy Particle Swarm Optimization", Proceedings of the second Joint Congress on Fuzzy and Intelligent Systems, Malek Ashtar University of Technology, Tehran, Iran, 28-30 October, 2008

[13] B. Niu and Y. Zhu, X. He, H. Wu, "MCPSO: A multi-swarm cooperative particle swarm optimizer", Applied Mathematics and Computation 185, Elsevier, 2007, pp. 1050–1062.

[14] B. Niu and Y. Zhu and X. Xian and H. Shen, "A multi-swarm optimizer based fuzzy modeling approach for dynamic systems processing ", Elsevier, 2007.

[15] F. V. D. Bergh and A. Engelbrecht, "A cooperative approach to particle swarm optimization", IEEE Transactions on Evolutionary Computation, Vol. 8, No. 3, 2004.

[16] F. V. D. Bergh and A. Engelbrecht, “Cooperative learning in neural networks using particle swarm optimizers”, South African Comput., vol. 26, 2000, pp. 84–90.

[17] B. Jiang, “Spatial clustering for mining knowledge in support of generalization processes in GIS”, ICA Workshop on Generalisation and Multiple representation, 2004, Leicester.

[18] J. Vesanto and Esa Alhoniemi, “Clustering of the self-organizing map”, IEEE Transactions on Neural Networks, Vol. 11, No. 3, pp. 586-600, 2000.

[19] X.Cui and E. Thomas, “Document clustering using particle swarm optimization”, Oak Ridge, IEEE, 2005.

[20] S. Selim and M. Ismail, “K-means type algorithms: A generalized convergence theorem and characterization of local optimality”, IEEE, 1984.

[21] Maytal Saar-Tsechansky, Foster Provost, “Handling Missing Values when Applying Classification Models”, Journal of Machine Learning Research 8 (2007) 1625-1657

[22] C.K. Goh, K.C. Tan, D.S. Liu, S.C. Chiam , “A competitive and cooperative co-evolutionary approach to multi-objective particle swarm optimization algorithm design” European Journal of Operational Research, Volume 202, Issue 1, 1 April 2010, Pages 42-54.

[23] Jiuzhong Zhang, Xueming “A Multi-Swarm Self-Adaptive and Cooperative Particle Swarm Optimization” Engineering Applications of Artificial Intelligence, Volume 24, Issue 6, September 2011, Pages 958-967