Study on An improvement of Numerical Association Rule Extraction for Multi-Objective Optimization Problem (Case studi: Bioelectric Potential Data)

Imam Tahyudin, Hidetaka Nambo (Chief Supervisor)
Student ID : 152 404 2008
Artificial Intelligence Laboratory,
Graduate School of Natural Science and Technology
Division of Electrical Engineering and Computer Science,
Kanazawa University, Japan
imam@blitz.ec.t.kanazawa-u.ac.jp

June 28, 2018

Abstract

PSO for solving the numerical association rule mining (ARM) problem has some weakness. Among of them is premature to search the optimal solution because it traps in local solution. This research proposed a method to overcome that problem by combining PSO method with Cauchy distribution (PARCD method). The main purpose is to develop PSO method in numerical ARM problem and to design, implement, and evaluate the method for bioelectric potential data set. The result showed that PARCD method has promise result.

Furthermore, another problem is the accuracy for estimating the human position around plant of bioelectric potential. The previous researches have been conducted by using some methods, such as decision tree (J48), multi layer perceptron (MLP), deep learning (CNN), and etc. Those accuracy methods still under 60%. Therefore, we proposed the different approach using association analysis method. After we got the best rules using association analysis method, we did matching process to calculate how many numbers of rules which precise. Finally, we got the best number for estimating the human position with the accuracy is around 75%.

Moreover, we proposed another method by using time series approach. And then, we got the best model is seasonal ARIMA model (1,0,0) with the accuracy is around 80%.
Keyword: Association rule mining, PSO, MOPAR, PARCD, bioelectric potential of plant, time series, SARIMA, estimation position

1 Improved Optimization of Numerical Association Rule Mining by Hybrid PSO and Cauchy Distribution Approach

1.1 Introduction

The ARM or association analysis method is used to find associations or relationships between variables, which often arise simultaneously in a dataset [1]. The Apriori and Frequent Pattern (FP) growth methods are widely employed in association analysis. These methods are suitable for categorical or binary data, such as gender data [2]. In addition, both methods require manual intervention to determine the minimum support (attribute coverage) and confidence (accuracy) values [3], [4]. To resolve this problem, some researchers have proposed solutions that employ optimization approaches without determining the minimum support and minimum confidence. However, this method can also become trapped in local optima.

We proposed a method that can address the premature searching and the limitations of traditional methods that it does not use a discretization process. One solution is by combining PSO with the Cauchy distribution (PARCD) method.[5].

1.2 PSO for Numerical Association Rule Mining with Cauchy Distribution

PARCD is an extension of the MOPAR methods that combines PSO and the Cauchy distribution to solve problems that occur in the association analysis of numerical data [6]. The velocity function is expressed as follows,

$$\begin{aligned} V_i(t+1) =& \omega(t)V_i(t) + C_1 rand()(pBest - X_i(t)) + \\ & C_2 rand()(gBest - X_i(t)) \end{aligned} \tag{1}$$

The next step is normalization by using $V_i(t+1)$ value (1),

$$U_i(t+1) = \frac{V_i(t+1)}{\sqrt{V_{i1}(t+1)^2 + V_{i2}(t+1)^2 ... + V_{iK}(t+1)^2}} \tag{2}$$

The result of the normalization process is multiplied by the Cauchy random variable as follows.

$$S_i(t+1) = U_i(t+1) \cdot tan\left(\frac{\pi}{2} \cdot rand[0,1)\right) \tag{3}$$

Then, the result of Eq. 3 which is a combination of the velocity value and the Cauchy distribution, is used to determine the new position of a particle.

$$X_i(t+1) = X_i(t) + S_i(t+1) \tag{4}$$

1.3 Result and Discussion

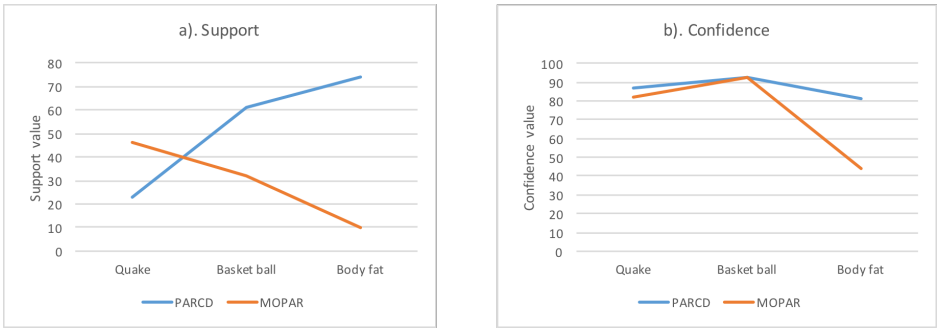

Figure 1: Comparison of support and confidence function

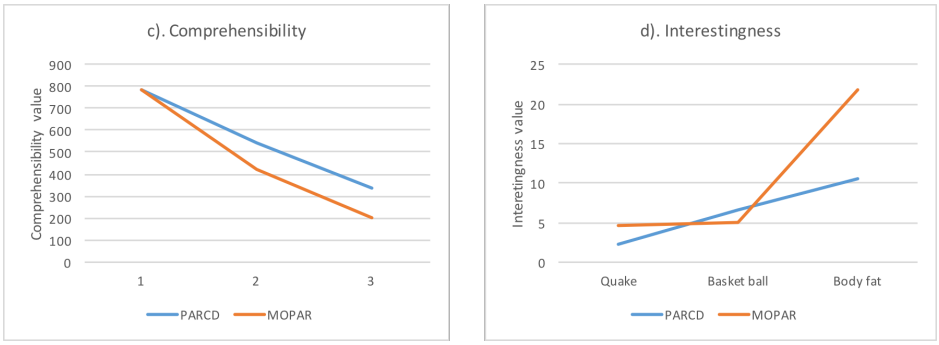

Figure 2: Comparison of comprehensibility and amplitude function

According to Fig. 1 and Fig. 2, the overall results indicate that proposed PARCD method obtained multi-objective function is better compared to the existing methods when searching for an optimal value.

1.4 Summary

This study has proved that combining the PSO with Cauchy distribution can solve the numerical ARM problem. The problems of local minimum and premature convergence with large datasets can be solved using the proposed

method. The experimental results demonstrate that the proposes PARCD method outperforms existing methods to all multi-objective functions, such as the support, confidence, comprehensibility, and interestingness.

2 Bioelectric Potential Plant for Determining Human Position

2.1 Introduction

Based on the results from previous studies, the bioelectric potential of plants has the ability to capture human behavior well. Research conducted by Shimbo et al. has shown that human behavior such as touching the plant, opening the door, approaching the plant, and turning on the lighting can be detected by extracting the characteristic of bioelectrical potential of plants [7]. Subsequent research conducted by Jin et al. used an artificial neural network algorithm to successfully detect the distance of person from a plant by observing the plants bioelectric potential [8]. Another study conducted by Nambo et al. used the bioelectric potential of plants to estimate people were in a room. They used several algorithms including a decision tree (J48) for classification and a multilayer perceptron to determine the presence of people. Next, they carried out a regression model for a matching process. The results showed that a person's presence in a room could be determined with an accuracy rate of 60% [9] [10], [11]. These previous researches did not determine the specific position of people in the space. Therefore, this research aimed to estimate the exact position of people using the bioelectric potential of plants by association rule mining (ARM) with particle swarm optimization (PSO).

2.2 MOPAR Method

The MOPAR method used PSO for solving numerical association rule mining problem. There are two main formula for this method which given in Eq. 5 and Eq. 6. The first equation is the velocity model and the second one is the position model [12].

$$V_i^{new} = \omega V_i^{old} + C_1 rand()(pBest - X_i) + C_2 rand()(gBest - X_i) \quad (5)$$

$$X_i^{new} = X_i^{old} + V_i^{new} \quad (6)$$

Here ω is the inertia weight; V_i^{old} is the velocity of the ith particle before updating; V_i^{new} is the velocity of the ith particle after updating; X_i is the ith,

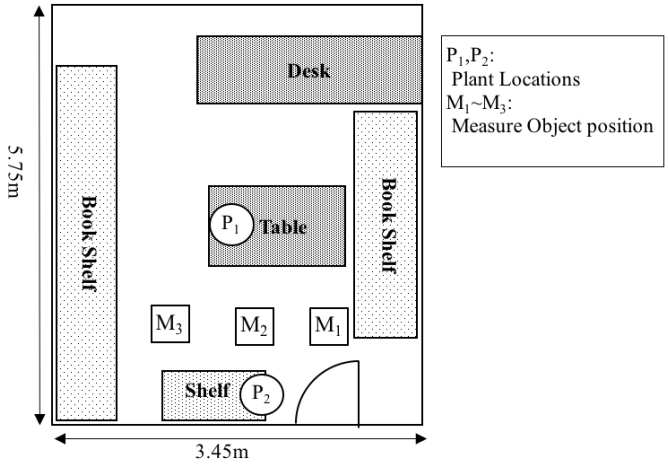

Figure 3: Experimental environment

or current particle; i is the number of particles; $rand()$ is a random number in the range (0, 1); C_1 is the cognitive component; C_2 is the social component; $pBest$ is the particle best or local optima in some iterations on every running; $gBest$ is the global best or global optima in some iterations on every running. Particle velocities in each dimension are restricted to maximum velocity V_{max} [12], [13].

2.3 Experiments and Results

2.3.1 Bioelectic Potential of Data set

The experiment was conducted in a room 5.75 m x 3.45 m with three object positions M_1, M_2, and M_3 and two plants, P_1 and P_2 (Fig. 3). A person walked around each object for 30 second. Two plants detect a responded to that human action, and the changes in the data were seen on the monitor of the data logger. The results of the spectrum were recorded and saved in the PC.

2.3.2 Matching Process and Evaluation

This process has as aim to know what the position of the human who walks around the object if there is a new data set (a testing data set). The testing data set which appropriate with one of the rules is indicated that person is found around that object position. The evaluation resulted in almost 75 % accuracy.

Table 1: Number of Matching Rules

	Position 1	Position 2	Position 3
Rule 1	1	1	0
Rule 2	2	11	6
Total	3	12	6

2.4 Summary

The purpose of this research was to determine the position of people using the bioelectric potential of plants. This has been successfully achieved. Using association analysis and a PSO approach; MOPAR has performed the estimation of the location of a person in one of three positions with an accuracy of approximately 75%. In the future, we will improve the accuracy by other methods and other approaches such as by applying a modified MOPAR method, time series model, or deep learning analysis.

3 SARIMA model for Bioelectric potential of plant

3.1 Introduction

The use of time series for bioelectric potential data set has been conducted using some models which are autoregressive (AR), moving average (MA), AR with grid search optimization, and ARMA model. Their average of forecasting accuracy were around 75% [6], [14], [15].

Accroding to the previous research, time series has robust ability to estimate and to predict some cases in many fields even when it combines with other methods. Therefore, this reasearch is intereseted for solving bioelectric potential of data set that to obtain the best model.

3.2 Proposed Method

3.2.1 ARIMA Model

ARIMA is a time series model which consist of Autoregressive (AR), Integration(I), and Moving average (MA). Generally there are two kinds of ARIMA model which are ARIMA non seasonal and ARIMA seasonal. This proposed method uses the seasonal ARIMA type. The ARIMA model is writen as

ARIMA(p,d,q) which p is the number of AR term, d is the number of I, and q is the number of MA term. The general model of ARIMA(p,d,q) is see in eq. 7 [16].

$$(1 - \phi_1 B.... - \phi_p B^p)(1 - B)^d Y_t = c + (1 + \theta_1 B..... + \theta_q B^q)e_t \tag{7}$$

There are three main components, which are the first is AR(p) term,

$$(1 - \phi_1 B.... - \phi_p B^p) \tag{8}$$

the second is integration for differentiation (d),

$$(1 - B)^d Y_t \tag{9}$$

and the third is MA(q) term,

$$(1 + \theta_1 B..... + \theta_q B^q)e_t \tag{10}$$

In addition, c is a constant value.

3.3 Results and Discussion

3.3.1 SARIMA Model

For constructing the SARIMA model, the first step is by confirming the dataset visualisation. The result shows that the dataset performs stationary because mean and variance of bielectric potential data set runs constantly. After that, we check the visualization of ACF dan PACF value.

According to acf plot we detected that the data set is seasonal because every eight leg has the same pattern. Therefore, we sholud get rid this seasonality using differentiation. By using R software, we obtain the SARIMA model (1,0,0) from 15.000 instances. It means the order of seasonal AR is 1, integrated and MA orders are zero.

SARIMA model (1,0,0) is the best model which is obtained by comparing among the other models because the AIC and log likelihood is the less one. In other word, It can be explained clearly that SARIMA (1,0,0) consists p=1, d=0 and q=0. According to the result the coefficient and intercept values are 0.9374 and -0.6997 respectively. Furthermore, the AIC value is -56285.83 and log likelihood value is 28146.92. For make sure that this model is better than others by calculating the accuracy. This is conducted using bioelectric potential dataset which is devided into 30 groups. Each group is 500 instances and we obtained the best SARIMA model from all groups.

Table 2: Testing result

SARIMA model	Accuracy
(2,0,0)	3.3%
(3,0,0)	16.7%
(1,0,0)	80%

After that, we calculate the accuracy by calculating the proportion. The result is presented in the table 2.

This table compare three SARIMA model that SARIMA (1,0,0) is the best model because it has the highest accuracy of 80%. The remain SARIMA model, SARIMA (3,0,0) and SARIMA (2,0,0), the accuracy are 16,7% and 3.3% respectively.

3.4 Summary

Time series approach has the contribution to enhance bioelectric potential of plant study. The data has appropriate characteristic to analyze. Because of the data set is detected has seasonality, we use SARIMA model. Finally, SARIMA model (1,0,0) is the best model for this research. This model has AIC value of -56285.83 and accuracy of 80% . In addition, the architecture design of bioelectric potential of plant for estimating human position is interested to follow for next research.

Conclusion and Future Work

The PARCD method that we proposed and implemented in this research is competitive and robust for solving the numerical association rule mining problem. Furthermore, this method could estimate the human position estimation using bioelectric potential of plant as well with the accuracy almost 75%. In addition, the seasonal ARIMA model (1,0,0) for the time series approach is the best model with the accuracy of this model is about 80%. For the future research, the improvement of optimization method for numerical ARM is still open. We can try by other combination such as with GA, fuzzy logic, deep learning, and etc. In addition, for time series approach will use SARIMA model for determine human position and also the model will be optimized by using evolutionary algorithms.

References

[1] H. Jiawei, K. Micheline, and P. Jian, *DATA MINING (Concept and Techniques)*, 2012, vol. 3, no. 13.

[2] I. H. Witten, F. Eibe, and H. Mark A., *Data Mining (Practical Machine Learning Tools and Techniques)*. Elsevier, 2011, vol. 3, no. 9.

[3] X. Yan, C. Zhang, and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3066–3076, mar 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408000195

[4] H. R. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 288–298, jan 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410005646

[5] M. Gen; L. Lin; H.Owada, "Hybrid Evolutionary Algorithms and Data Mining: Case Studies of Clustering," *Proc. Soc. Plant Eng. Japan 2015 Autumn Conf.*, 2015.

[6] I. Tahyudin and H. Nambo, "The Combination of Evolutionary Algorithm Method for Numerical Association Rule Mining Optimization," in *Tenth Int. Conf. Manag. Sci. Eng. Manag.* Baku, Azerbaijan: Springer Berlin Heidelberg, 2016, p. 1. [Online]. Available: http://www.icmsem.org/

[7] K. Nomura, H. Nambo, and H. Kimura, "Development of Basic Human Behaviors Cognitive System using Plant Bioelectric Potential," *IEEJ Trans. Sens. Micromachines*, vol. 134, no. 7, pp. 206–211, 2014.

[8] X. Jin, "Recognition of the Distance between Plant and Human by Plant Bioelectric Potential," *APIEMS*, pp. 602–606, 2014.

[9] H. Nambo, "A Study on the Estimation Method of the Resident ' s Location using the Plant Bioelectric Potential," *APIEMS*, pp. 1896–1900, 2015.

[10] H. Nambo and H. Kimura, "Estimation of Resident ' s Location in Indoor Environment Using Bioelectric Potential of Living Plants," *Sens. Mater.*, vol. 28, no. 4, pp. 369–378, 2016.

[11] ——, "Development of the Estimation Method of Resident ' s Location using Bioelectric Potential of Living Plants and Knowledge of Indoor Bookshelf," in *Tenth Int. Conf. Manag. Sci. Eng. Manag.*, 2017.

[12] V. Beiranvand, M. Mobasher-Kashani, and A. Abu Bakar, "Multi-objective PSO algorithm for mining numerical association rules without a priori discretization," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4259–4273, jul 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414000025

[13] Y. Xinjie and M. Gen, *Introduction to Evolutionary Algorithms*. Springer London Dordrecht Heidelberg New York, 2010.

[14] I. Tahyudin and H. Nambo, "Comparison Study of Time Series Model on Bioelectric Potential Data set," in *SOPEJ*, 2017, pp. 1–5.

[15] ——, "An optimization of Numerical Association Rule Mining by Using A combination of PSO and Cauchy Distribution," -.

[16] J. Hamilton, *Time Series Analysis*. Princeton University Press, Princeton, New Jersey., 1994.

平成30年　8月　1日

学位論文審査報告書（甲）

1．学位論文題目（外国語の場合は和訳を付けること。）

STUDY ON AN IMPROVEMENT OF NUMERICAL ASSOCIATION RULE EXTRACTION FOR MULTI-OBJECTIVE OPTIMIZATION PROBLEM (Case Study:Bioelectric Potential of Plant Data)

（多目的最適化における数値属性相関ルール抽出法の改善に関する研究 -植物生体電位データへの応用）

2．論文提出者　(1) 所　　属　電子情報科学専攻

(2) 氏（ふり）名（がな）　IMAM TAHYUDIN（イマム タヒュディン）

3．審査結果の要旨（600～650字）

平成30年8月1日に第1回学位論文審査委員会を開催し、同日口頭発表を実施した。その後、引き続き第2回学位論文審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

数値属性を持つ相関ルールを効率的に最適化する手法として、遺伝的アルゴリズム(Genetic Algorithm:以下GA)や粒子群最適化法(Particle Swarm Optimization:以下PSO)に代表されるメタヒューリスティクスを用いた手法が挙げられる。本論文では、PSOを用いた数値属性相関ルールの最適化手法に着目した。PSOを用いる場合、最適解を導く過程で局所解に陥るという問題に対して様々な改善手法が提案されており、これまで粒子の速度更新式にGAや正規分布を導入する手法が提案されている。本論文では、粒子の速度更新式にコーシー分布に基づく確率項を導入することで、従来法と比較して適応度の高い相関ルールを導く手法を提案し、ベンチマークデータを用いてその効果を検証した。さらに、実証実験として、植物生体電位による居住者の位置推定問題に提案手法を適用し、得られた相関ルールを用いることで、決定木や多層パーセプトロンを用いた従来手法よりも高い精度での位置推定を実現した。

以上のように、本研究は数値属性相関ルールの最適化に関して効果的な手法を提案しており、データマイニングの一分野の発展に貢献するものである。よって、博士（工学）に値すると判定した。

4．審査結果　(1) 判　　定（いずれかに○印）　○合　格　・　不合格

(2) 授与学位　　博　士（　工学　）