A METHOD FOR FACTOR SCREENING OF SIMULATION EXPERIMENTS BASED ON ASSOCIATION RULE MINING

Lingyun Lu[a], Wei Li[b], Ping Ma[c], Ming Yang[d]

Control and Simulation Center, Harbin Institute of Technology, Harbin 150080, P.R. China

[a]jyzluhit@gmail.com, [b]frank@hit.edu.cn, [c]pingma@hit.edu.cn, [d]myang@hit.edu.cn

ABSTRACT

Since complex simulation models contain a large number of inputs and parameters, it is meaningful to identify the most important ones for further studies. The traditional procedure of factor screening is to design simulation experiment schemes and generate simulation output data first. Furthermore, due to rigorous hypothesis and fewer levels for each factor, the existing screening methods are usually not appropriate for complex simulation. In this paper, a new method for factor screening of simulation experiments is proposed. It is based on data mining technique and existing simulation data. The relationships between factors and the output are discovered from the sets of simulation data using the association rule mining algorithm firstly. Then the important factors can be identified by synthesizing the association rules. Finally, the proposed method is verified by the test function.

Keywords: simulation experiments, factor screening, association rules, quantitative attributes

1. INTRODUCTION

Complex simulation models consist of a large number of inputs and parameters which are generally referred to as factors in design of experiments (DOE). DOE is essential for doing certain analyses include validation, uncertainty analysis, sensitivity analysis, optimization, etc (Kleijnen 2008, Li 2016). However, it is usually prohibitive or impractical to do these analyses with the large number of factors involved, especially for computer simulation models with high computational cost. An important question is – which factors are really significant when there are potentially a large number factors involved (Alam 2004)? The Pareto principle or 20-80 rule implies that only a few factors are really important (Wright 2010). In order to improve efficiency, it is reasonable to screen a small number of factors before analyzed. Factor screening is performed to eliminate unimportant factors so that the remaining important factors can be more thoroughly studied in later experiments, regardless of the application domain of area, including national defense, logistics, industry, healthcare, etc (Li 2016, Bruzzone 2014, Padhi 2013, Rantanen 2015).

Several types of factor screening methods have been developed to identify important factors with a limited set of simulation experiments. The most common ones are fractional factorial, central composite, and Plackett–Burman designs (Myers 2002). However, these screening designs developed for physical experiments are not appropriate for simulation physical experiments. Procedures developed for stochastic simulation experiments include Morris' randomized one-factor-at-a-time design (Morris 1991), frequency domain method (Morrice 1995), edge designs (Elster 1995), iterated fractional factorial designs (Campolongo 2000) and the Trocine screening procedure (Trocine 2001). Some methods developed for discrete-event simulation experiments include sequential bifurcation (SB), SB with interactions (SB-X), Cheng's method, controlled sequential bifurcation (CSB), CSB with interactions (CSB-X), controlled sequential factorial design (CSFD), fractional factorial controlled sequential bifurcation (FFCSB), etc (Wan 2007, Shen 2009). However, the majority of these methods have many assumptions or limitations, such as monotonicity of the function, the smoothness of the inputs/outputs function, etc. These methods are efficient and effective when their assumptions are satisfied.

From the factor screening methods described above, it can be concluded that these methods easily neglect some important factors due to rigorous hypothesis and few factor levels. It is possibly unsuitable for the complex simulation models. As is known to all, the importance of factors depends on the experimental domain. The effective way is to select levels as many as possible during factor screening which can represent simulation model more precisely and improve the accuracy of the classification. So the data mining algorithm is adopted to solve this problem, which can be based on the existing simulation data and regardless of mathematics assumptions.

In this paper, the association rule mining algorithm is investigated as a potential factor screening method. It should be based on simulation data, including simulation inputs/parameters and the output. The remainder of the paper is organized as follows. In Section 2, the procedure of the Apriori algorithm and the algorithm of partitioning quantitative attributes are described. In Section 3, a method for factor screening of simulation experiments is proposed. Through this method, important factors can be identified by synthesizing the association rules. In Section 4, the proposed method is illustrated and verified by the test function. Finally, the paper is concluded in Section 5.

2. RELATED WORK

The objective of factor screening is to identify the important factors among many factors involved. In this section, the procedure of the Apriori algorithm will be described. However, this algorithm can only cope with boolean attributes instead of quantitative attributes. But quantitative attributes usually exist in simulation models. So the algorithm of partitioning quantitative attributes is adopted.

2.1. Apriori Algorithm

The Apriori algorithm is proposed by Agrawal (1994) which is used to find potential relationships between the items (Data_Attributes). The result of the algorithm will be expressed in the form of association rules. In order to clarify the Apriori algorithm clearly, some notations is illustrated in the following.

- *k*-itemset: An itemset having k items.
- L_k : Set of large *k*-itemsets with minimum support.
- C_k : Set of candidate *k*-itemsets which has potentially large itemsets.
- D : Set of transactions where each transaction T is a set of items, i.e., $T \subseteq D$.

Given a set of transactions D , the use of the Apriori algorithm is to generate all association rules which the support and confidence are larger than the given minimum values. The minimum support and minmum confidence are referred to as *min_sup* and *min_conf* respectively. The procedure of the Apriori algorithm is described as follows:

```
Apriori algorithm
$L_1$={large 1-itemsets};
For ( $k=2$ ; $L_{k-1} \neq \varnothing$ ; $k++$ ) do begin
    $C_k$ = apriori-gen( $L_{k-1}$ );
    For all transactions $t \in D$ do begin
        $C_t$ = subset( $C_k$ , $t$ );
        For all candidates $c \in C_t$ do
            $c$.count++ ;
        end
    end
    $L_k = \{ c \in C_k \mid c.\text{count} \geq min\_sup \}$;
end
return $L = \quad _k L_k$ .
```

At the beginning, the large 1-itemsets is determined by simply counts item occurrences. Then subsequent pass k consists of two phases. One is the candidate itemsets C_k which is generated by the apriori-gen function; and the other is the support of candidates in C_k which is counted by scanning the database.

The apriori-gen function described above returns a superset of the set of all large *k*-itemsets. And the subset function returns all the candidates contained in a transaction t with some rules. The more details of these two functions can be referred to the literature Agrawal (1994).

2.2. The Algorithm of Partitioning Quantitative Attributes

However, quantitative attributes need to be discretized when using the Apriori algorithm. Two problems may be occurred after discretization (Srikant 1996).

- ***min_sup***. If the number of intervals for a quantitative attribute (or values, if the attribute is not partitioned) is large, the support for any single interval can be low. Hence, without using larger intervals, some rules involving this attribute may not be found because they lack minimum support.
- ***min_conf***. There is some information lost whenever we partition values into intervals. Some rules may have minimum confidence only when an item in the antecedent consists of a single value (or a small interval). This information loss increases as the interval sizes become larger.

In order to solve problems described above, the algorithm of the equi-depth partitioning (Fukuda 1999) is adopted, which can divide N data into M buckets almost evenly. The procedure of the algorithm is described as follows.

- **Step 1**: Make an *S*-sized random sample from N data.
- **Step 2**: Sort the sample in $O(S \log S)$ time.
- **Step 3**: Scan the sorted sample and set $i(S/M)$ -th the smallest sample to p_i for each $i=1, \quad ,M-1$. Let p_0 be $-\infty$ and p_M be $+\infty$.
- **Step 4**: For each tuple x in the original N data, find i such that $p_{i-1} < x \leq p_i$ and assign x to the *i*-th bucket. This check can be done in $O(\log M)$ time by using the binary search tree for the buckets. Thus, for all i , the size u_i of B_i can be computed in $O(N \log M)$ time.

3. THE METHOD FOR FACTOR SCREENING OF SIMULATION EXPERIMENTS

To identify important factors from many factors involved in simulation models, the Apriori algorithm described in Section 2 is adopted. However, in many cases, factors used in simulation are quantitative attributes, which the Apriori algorithm cannot cope with them. So data preprocessing is used to partition the domain of quantitative attributes into several intervals and mapped these intervals into discrete attributes. And the algorithm of the equi-depth partitioning is adopted to pruning the search space. What need to be pointed out that calculation of strong rules is not interesting while using support-confident framework, lift for mining positive non-redundant association rules is adopted here.

It is assume that itemsets A and B are denote the premise and consequent of association rules, respectively, where $A \neq \varnothing$, $B \neq \varnothing$, $A \quad B = \varnothing$. Here $A \Rightarrow B$ is used to represent the relationship between

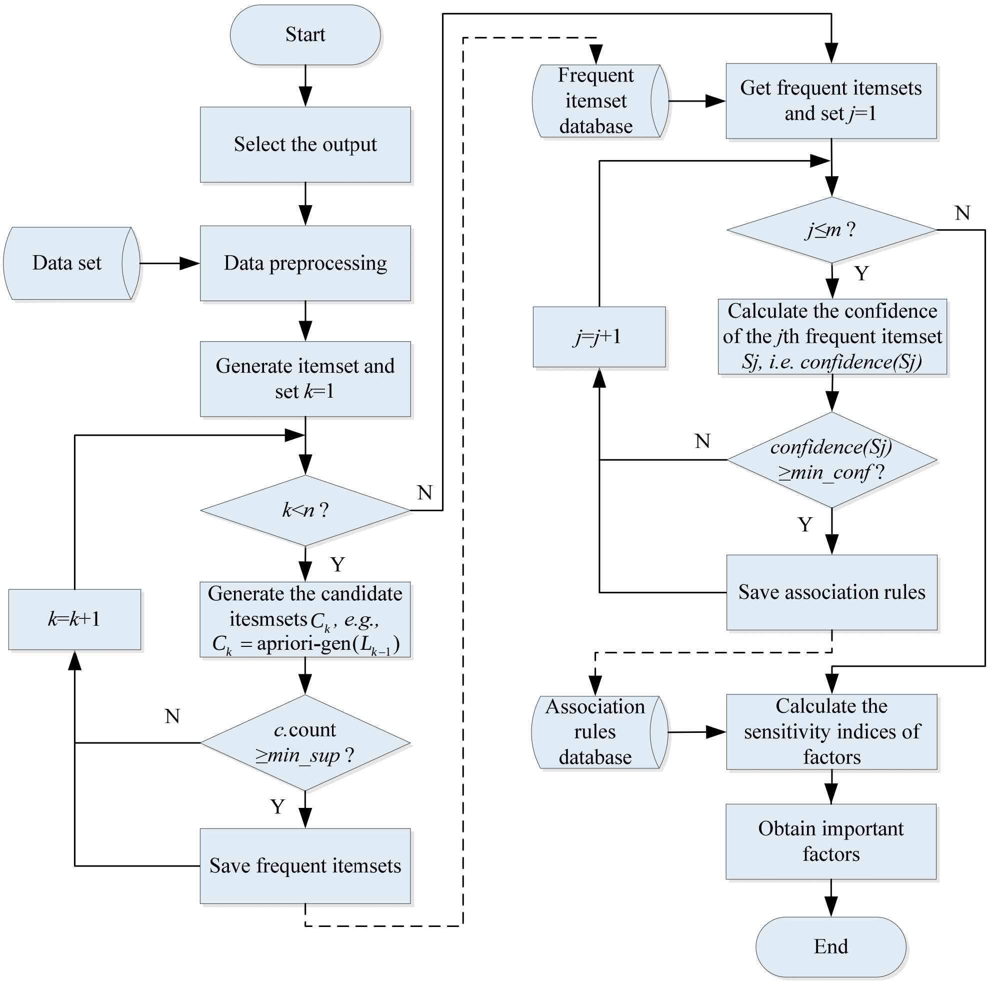
Start
Select the output
Data set
Data preprocessing
Generate itemset and set k=1
k<n ?
N
Y
Generate the candidate itesmsets C_k, e.g., C_k = apriori-gen(L_{k-1})
k=k+1
N
c.count ≥min_sup ?
Y
Save frequent itemsets
Frequent itemset database
Get frequent itemsets and set j=1
j≤m ?
N
Y
Calculate the confidence of the jth frequent itemset Sj, i.e. confidence(Sj)
j=j+1
N
confidence(Sj) ≥min_conf ?
Y
Save association rules
Association rules database
Calculate the sensitivity indices of factors
Obtain important factors
End

Figure 1: Flow of Factor Screening Based on the Improved Apriori Algorithm

factors and the output. Then the confidence of $A \Rightarrow B$ can be calculated as

$$confidence(A \Rightarrow B) = P(B \mid A) = \frac{support(A \quad B)}{support(A)} \qquad (1)$$

where $support(A)$ is the support of A.

The flow of factor screening based on the improved Apriori algorithm is shown in Figure 1. The procedure of the proposed method mainly consists of four steps:

- **Step 1**: Data preprocessing consists of partition quantitative attributes into several intervals and discretization.
- **Step 2**: Obtain the set of all frequent itemsets by finding all itemsets where the support of the corresponding itemsets satisfies *min_sup*.
- **Step 3**: Generate association rules using frequent iemsets where the confidence of the corresponding frequent itemsets satisfies *min_conf*.
- **Step 4**: Calculate the sensitivity indices of factors by Equation (2), and eventually obtain the important factors by comparing the values of sensitivity indices.

The equation of calculating sensitivity indices is given as

$$\begin{aligned} r_i &= \max(\text{confidence}(\text{x}_i \Rightarrow \text{y})) - \\ &\quad \min(\text{confidence}(\text{x}_i \Rightarrow \text{y})) \qquad (2) \\ S_i &= \frac{r_i}{\sum r_i} \end{aligned}$$

where x_i is factor and y is the output.

4. CASE STUDY

The method proposed in this paper can be used for factor screening of simulation experiment. In this section, a test function is used to demonstrate the effectiveness of the proposed method.

Consider the test function

$$y = 11x_1 + 9x_2 + 2x_1x_2 + 15x_3 + 3x_4 - 2x_5 + 16x_6 + x_7 \qquad (3)$$

where $x_1, x_2, \quad , x_7$ is an i.i.d. sample from $\boldsymbol{U}(0,1)$.

The set of data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_7, \boldsymbol{y}\}$ is obtained first, where the sample size is 5000. Then each factor x_i is partitioned into seven intervals through the algorithm of the equi-depth partitioning. Each interval of x_i and y are discretized in the form of $\{x_i\_1, x_i\_2, \quad , x_i\_7\}$ and $\{y\_1, y\_2, \quad , y\_7\}$, respectively. The association

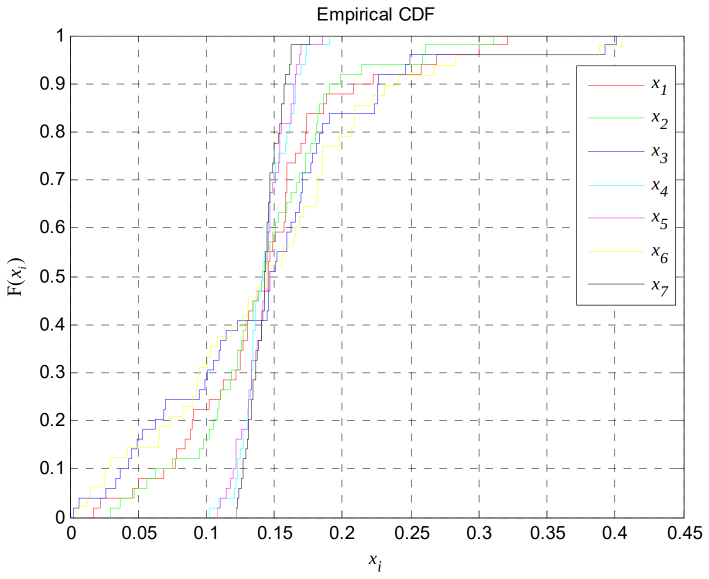

Figure 2: An Ensemble of 7 CDFs of the Factors

rules are generated by the Apriori algorithm, as shown in Table 1. Only 2-itemsets is taken to be considered, i.e., A only contains one factor in each rule. According to the fourth column of Table 1, the cumulative distribution function (CDF) for the confidence of each factor x_i is drawn in Figure 2. The relative importance of the factors is qualitatively expressed by the gradients of the curves, which the smaller one is more important. In other words, the range of confidence for each factor reflects the importance of factors. Furthermore, the sensitivity indices for each factor are calculated and shown in Figure 3. It can be quantitatively conclude that the order of factors is $x_6 > x_3 > x_1 > x_2 > x_4 > x_5 > x_7$ according to the importance. The important factors are x_6, x_3, x_1 and x_2. The experimental results obtained above are in agreement with the theoretical results which are easily concluded from Equation (3).

Table 1: List of the Association Rules

ID	A	B	$confidence(A \Rightarrow B)$
1	x_1_2	y_2	0.1737
2	x_1_2	y_5	0.1275
3	x_1_4	y_6	0.1569
137	x_3_1	y_5	0.0700
249	x_6_7	y_7	0.4050
342	x_7_3	y_4	0.1261

Above all, it is demonstrated that the proposed method in this paper is an effective method which can be used for factor screening. Compared with the traditional screening methods, the proposed method can identify important factors based on the existing simulation data and regardless of mathematics assumptions.

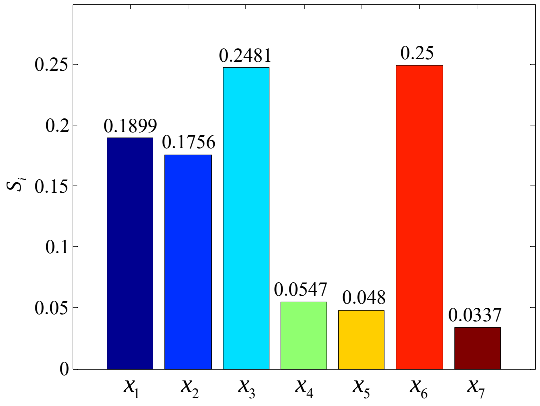

Figure 3: Histogram of Sensitivity Indices for Test Function Using the Proposed Method

5. CONCLUSIONS

Factor screening is very important while simulation models to be analyzed with many factors involved. In this paper, the use of association rule mining algorithm as a kind of factor screening method has been studied. The relationships between factors and the output are considered to be a measure of index to identify important factors from many factors involved. As a classic algorithm of association rule mining, Apriori algorithm can only be used to solve the problem of mining boolean association rules. So the algorithm of equi-depth partitioning is adopted to transform quantitative attributes into boolean attributes. After the association rules are generated, the results of confidence are used to calculate the sensitivity indices of factors. And the important factors are identified according to the sensitivity indices. It is demonstrated from the case study that the proposed method appears to perform well in factor screening.

To improve the efficiency of factor screening and insure its accuracy, two aspects are deserved to be further studies. One is to study how to synthesize the association rules in other ways, and the other is how to process the unbalanced data.

ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (Grant No. 61403097).

REFERENCES

Kleijnen J.P.C., 2008. Design and analysis of simulation experiments. New York:Springer.

Li W., Lu L.Y., Liu Z.Z. et al., 2016. HIT-SEDAES: An integrated software environment of simulation experiment design, analysis and evaluation. International Journal of Modeling, Simulation, and Scientific Computing, 7 (3), 1650027.

Alam F.M., McNaught K.R., Ringrose T.J., 2004. Using Morris' randomized OAT design as a factor screening method for developing simulation metamodels. Proceedings of the 2004 Winter Simulation Conference, pp. 949-957. December 5-8, Washington DC, United States.

Wright A., Bates D.W., 2010. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at work. Applied Clinical Informatics, 1 (1), 32–37.

Sanchez S.M., Wan H., 2015. Work smarter, not harder: a tutorial on designing and conducting simulation experiments. Proceedings of the 2015 Winter Simulation Conference, pp. 1795-1809. December 6-9, Huntington Beach, United States.

Bruzzone, A., Longo, F., 2014. An application methodology for logistics and transportation scenarios analysis and comparison within the retail supply chain. European Journal of Industrial Engineering, 8 (1), 112–142.

Padhi S.S., Wagner S.M., Niranjan T.T. et al., 2013. A simulation-based methodology to analyse production line disruptions. International Journal of Production Research, 51 (6), 1885–1897.

Rantanen J., Khinast J., 2015. The future of pharmaceutical manufacturing sciences. Journal of Pharmaceutical Sciences, 104 (11), 3612–3638.

Myers, R.H., Montgomery, D.C., 2002. Response surface methodology: process and product optimization using designed experiments, second ed. John Wiley & Sons, New York.

Morris M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics, 33: 161–174.

Morrice, D.J., Bardhan, I.R., 1995. A weighted least squares approach to computer simulation factor screening. Operations Research, 43 (5), 792–806.

Elster, C., Neumaier, A., 1995. Screening by conference designs. Biometrika 82 (3), 589–602.

Campolongo, F., Kleijnen, J.P.C., Andres, T., 2000. Screening methods. In: Saltelli, A., Chan, K., Scott, E.M. (Eds.), Sensitivity Analysis. John Wiley & Sons, New York, pp. 65–89.

Trocine L., Malone L.C., 2001. An overview of newer advanced screening methods for the initial phase in an experimental design. Proceedings of the 2001 Winter Simulation Conference, pp. 173-176. December 9-12, Arlington, United States.

Wan H., Ankenman B., 2007. Two-stage Controlled Fractional Factorial Screening for Simulation Experiments. Journal of Quality Technology, 39 (2), 126–139.

Shen H., Wan H., 2009. Controlled sequential factorial design for simulation factor screening. European Journal of Operational Research, 198, 511–519.

Agrawal R., Srikant R., 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499. September 12-15, Santiago, Chile.

Srikant R., Agrawal R., 1996. Mining quantitative association rules in large relational tables. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data pp. 1-12. June 4-6, Montreal, Canada.

Fukuda T., Morimoto Y., et al., 1999. Mining optimized association rules for numeric attributes. Journal of Computer and System Sciences, 58 (1), 1–12.

Rastogi R., Shim K., 2002. Mining optimized association rules with categorical and numeric attributes. IEEE Transactions on Knowledge and Data Engineering, 14 (1), 29–50.

Ke Y.P., Cheng J., Ng W., 2008. An information-theoretic approach to quantitative association rule mining. Knowledge and Information Systems, 16 (2), 213–244.

AUTHORS BIOGRAPHY

LINGYUN LU is a Ph.D. student at Harbin Institute of Technology (HIT), and received the M.E. from HIT in 2013. His research interests include simulation experiment design and simulation data analysis. His email is jyzluhit@gmail.com.

WEI LI is an associate professor at HIT, and received the B.S., M.E. and Ph.D. from HIT in 2003, 2006 and 2009 respectively. His research interests include simulation evaluation, simulation data analysis, and distributed simulation. His email is frank@hit.edu.cn.

PING MA is a professor at HIT, and received Ph.D. from HIT in 2003. Her research interests include distributed simulation technique and VV&A. Her email is pingma@hit.edu.cn.

MING YANG is the corresponding author. He is a professor and the director of control and simulation center at HIT. Also he is the vice editor-in-chief for Journal of System Simulation and editor for International Journal of Modeling Simulation and Science Computing. His research interests include system simulation theory and VV&A. His email is myang@hit.edu.cn.