Application of an Improved Association Rule Algorithm in Rural Development Assessment in China

Chenguang Zhang1, Guifa Teng2,*

1 School of Computer and Information Technology, Cangzhou Jiaotong College, Cangzhou, China; School of Information Science and Technology, Agricultural University of Hebei, Baoding, China
2 School of Information Science and Technology, Agricultural University of Hebei, Baoding, China

*Corresponding author e-mail: tguifa@126.com

Abstract. Association rules based on big data are derived from shopping basket analysis, which can deeply explore the association between things, and have been widely used in many fields. The traditional algorithm of association rules has several shortcomings, a new improved algorithm of association rules was proposed in this paper, which combines the precise advantages of Bayesian classifier with the full-scale characteristics of big data technology, and was applied to the rural development assessment process. After 2020, China's rural industrial structure will undergo great changes, and environmental protection to drive economic development is an important part of it. The new association rules can be better adapted to the needs of the assessment. Through comparison, it can be seen that the confidence results obtained by the new association rules are more accurate than Pearson correlation coefficient. Finally, the association rules among CPI, per capita consumption expenditure and per capita disposable income are analysed, and suggestions for rural development have also been made.

1. Introduction

Association rules [1] originated in shopping baskets and were first used to analyze shopping patterns, which can mine the inner relationship among things. It can be represented by $A \rightarrow B$ that $A$ is called the antecedent and $B$ is called the consequent. Support and confidence are its two units of measurement, like $\text{Support}(A) = \frac{|A|}{|N|}$ and $\text{Confidence}(A \rightarrow B) = \frac{|A,B|}{|A|}$. The final calculation result can usually be expressed as $\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$ in the form of Pearson correlation coefficient.

Apriori [2] is an early algorithm of association rules. Its principle is: for the same threshold, if a set of items is frequent, then every subset is also frequent [3]. There are many improved versions of this algorithm, such as FP-Growth [4], parallel algorithm based on Spark Streaming [5], heuristics algorithm [6], fuzzy algorithm [7] and so on, and it has been applied in many fields [8, 9, 10].

The development of rural is of great importance to a country [11]. China has made remarkable achievements in rural development and is the first developing country to achieve the Millennium Development Goals and halve the number of people living in poverty [12]. Compared with the early
days of reform and opening up 40 years ago, China's current rural overall environment and living standards have been greatly improved, and the number of people living in poverty has dropped from more than 500 million to 16.6 million in 2019 [13, 14]. At present, as the world's second largest economy, China's rural development is even somewhat resilient to the financial crisis [15].

After 2020, China's rural industrial structure will undergo tremendous changes, and environmental protection to drive economic development is an important part of it. In addition, a number of indicators will be watched, such as in the areas of education, health care and housing [16]. Traditional rural development evaluation methods have been unable to adapt to the actual needs, the new evaluation methods are very necessary.

2. The proposed algorithm of improved association rules

Take the rural related consumption data in Table 1 as an example. The ultimate purpose of the algorithm is to find the association rules among various indicators, so as to provide a reliable theoretical basis for the next stage of rural development policy formulation.

| Year | Food  | Clothing | Education | Medical | Housing |
|------|-------|----------|-----------|---------|---------|
| 2014 | 1830.62 | 358.37   | 450.95    | 537.19  | 1829.37 |
| 2015 | 1872.72 | 514.29   | 457.26    | 543.64  | 1743.56 |
| 2016 | 2157.77 | 576.63   | 560.15    | 634.29  | 2299.4  |
| 2017 | 2246    | 607      | 622       | 823     | 2045    |
| 2018 | 2528.16 | 701.17   | 730.8     | 957.94  | 2380.02 |

Data source: F County Economic Yearbook

In general, traditional association rule algorithms cannot solve such problem of evaluating indicators because the names and numbers of indicators are the same for each year, which makes it impossible for traditional algorithms to identify which indicator is more frequent. The improved association rules algorithm solves this problem by introducing Bayesian classifier.

Step1: determine support. In the process of indicator evaluation, any indicator may have potential value; therefore, each indicator must be retained. This means that the support is 1 and the pruning operation will not exist. Thus, the data in Table 1 can be represented as shown in Table 2.

| Year | Food  | Clothing | Education | Medical | Housing |
|------|-------|----------|-----------|---------|---------|
| 2014 | Food  | Clothing | Education | Medical | Housing |
| 2015 | Food  | Clothing | Education | Medical | Housing |
| 2016 | Food  | Clothing | Education | Medical | Housing |
| 2017 | Food  | Clothing | Education | Medical | Housing |
| 2018 | Food  | Clothing | Education | Medical | Housing |

Using the traditional big data algorithm (such as FP-Growth) to mine the association rules (the mining results are obtained from Mahout), the results are as follows in Table 3.

| Indicators | Frequent Item Sets |
|------------|--------------------|
| {Education} | ([Education], 5) |
| {Food}     | ([Education, Food], 5) |
| {Medical}  | ([Education, Food, Medical], 5) |
| {Clothing} | ([Education, Food, Medical, Clothing], 5) |
| {Housing}  | ([Education, Food, Medical, Clothing, Housing], 5) |
Obviously, the mining result of this association rule is not ideal; because in the actual economic development process of a region, it is very unlikely that all indicators have strong association rules at the same time. The reasons for this phenomenon are as follows.

The FP-Growth algorithm (other traditional association rule algorithms also face similar problems) is derived from the "shopping basket" analysis. The efficient execution of the algorithm should be based on two preconditions: First, the items (indicators) in each shopping basket (year) are different; Second, each item has a different number of purchases (occurrence times), so the item with the highest occurrence frequency can be selected. However, according to Table 1, there are five indicators (the same number) in every year, and the names (types) of the five indicators are also the same, which leads to errors in the algorithm due to the failure to select the most frequent item in the execution process. At this time, if the type or quantity of a certain indicator is adjusted, then it is not in accordance with the working method of the local poverty alleviation department, and the evaluation work will lose its basis.

Step2: increase the dimension. The purpose of poverty alleviation evaluation is to determine whether poverty can be lifted out in that year, and now poverty alleviation performance data is added as an auxiliary attribute, as shown in Table 4.

| Year | PCCE_l | I_1 | ... | I_m | A |
|------|--------|-----|-----|-----|---|
| Y_1  | CE_1   | X_{11}| ... | X_{1m} | A_1 |
| ...  | ...    | ... | ... | ... | ...    |
| Y_{m-1}| CE_{m-1}| X_{m-1,1}| ... | X_{m-1,m} | A_{m-1} |
| Y_m  | CE_m   | X_{m1} | ... | X_{mn} | A_m |

A represents the overall level of consumption in rural areas, the formula is defined as \( P(A) = \frac{\text{Income(now)}}{\text{Income(final)}} \), where \( \text{Income(now)} \) is annual per capita income current, and \( \text{Income(final)} \) is the level of annual per capita income after reaching a certain standard. That is to say, if the two are equal, then the value of performance is 1. PCCE is annual per capita consumption expenditure. As the order of magnitude difference between the indicators is large, the data need to be normalized. The normalization formula is divided into three steps: a) \( y_{ij} = \frac{x_{ij}}{CE_i} \), b) \( ce_i = CE_i / \sum_{j=1}^{n} CE_j \), c) \( z_{ij} = y_{ij} \times ce_i \).

Step3: Set \( Y_m \) the year of achieve a certain income level, and the rest are years of failure to achieve.

According to Bayes' formula \( P(x_{ij} \mid A) = \frac{P(A \mid x_{ij}) \times P(x_{ij})}{P(A)} \), we know that formula 
\[ P(A \mid x_{ij}) = \frac{P(x_{ij} \mid A) \times P(A)}{P(x_{ij})} \] holds. \( P(x_{ij} \mid A) \) is the data of not achieve, where \( 1 \leq i \leq m-1, 1 \leq j \leq n \).

According to the total probability formula, the existing probability of each indicator is \( P(x_{ij}) = P(x_{ij} \mid A)P(A) + P(x_{ij} \mid A)P(A) \), \( P(x_{ij} \mid A) \) is the data of achieve, where \( i = m, 1 \leq j \leq n \).

Step4: Using \( P(A \mid x_{ij}) = 1 - P(A \mid x_{ij}) \) to construct the matrix. It is the probability of achieving a predetermined income level, shown in Table 5. According to the result of probability matrix, the indicator with a high "per capita consumption expenditure" may not have a high "achieving certain income level probability", and the two are not in direct proportion. Therefore, statistical data cannot be used directly.

Step5: Constructing confidence matrix. According to the confidence formula \( \text{conf}(X \rightarrow Y) = \frac{\{X,Y\}}{\{X\}} \), the numerator is the number of occurrences of the item set \( \{X,Y\} \), namely the number of occurrences of \( X \) and \( Y \) together; The denominator represents the number of occurrences of the item set \( \{X\} \), that
is, the number of occurrences of $X$ alone. Therefore, confidence can be expressed by formula \( \text{conf}(X \rightarrow Y) = \frac{f(X,Y)}{N(X)} \).

**Table 5. Probability of Achieving Certain Income Level**

| Year | Food   | Clothing | Education | Medical | Housing |
|------|--------|----------|-----------|---------|---------|
| 2014 | 0.024  | 0.026    | 0.025     | 0.025   | 0.024   |
| 2015 | 0.027  | 0.027    | 0.028     | 0.028   | 0.027   |
| 2016 | 0.028  | 0.029    | 0.029     | 0.029   | 0.028   |
| 2017 | 0.029  | 0.029    | 0.029     | 0.029   | 0.030   |
| 2018 | 0.031  | 0.031    | 0.031     | 0.030   | 0.031   |
| 2019 | 0.032  | 0.032    | 0.032     | 0.032   | 0.032   |

Residents in this region do not communicate in advance about every consumption activity, so individual consumption activities are independent. When the statistical amount of personal consumption expenditure is large (the total rural population in the region), the confidence conforms to Bernoulli's law of large numbers \( \lim_{n \to \infty} P[|\frac{f(X,Y)}{N(X)} - P(X,Y)| < \varepsilon] = 1 \), where: \( \varepsilon \) is an arbitrarily small positive number. At this time, the solution of confidence is transformed from the quantity problem of frequent item set to the probability problem of indicator, which is equivalent to calculating the proportion of common part of $X,Y$ in $X$.

Therefore, the confidence formula of poverty alleviation evaluation can be set as \( \text{conf}(X \rightarrow Y) = \frac{\sum_{i=1}^{m} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{m} (X_i - \bar{X})(X_i - \bar{X})} \), where: the numerator represents the degree of correlation between any two indexes, and the denominator represents the degree of correlation between an index and itself.

When the constant \( \frac{1}{m-1} \) is contemporaneous in the numerator and denominator, the denominator becomes \( s^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})(X_i - \bar{X}) \), this is the sample variance and the unbiased estimate of the population variance. The numerator is \( \text{Cov}(X,Y) = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})(Y_i - \bar{Y}) \), it is the unbiased estimate of the sample covariance of the indicator. However, note that the denominator expands to its own square, so the square root operation is required. To make up for the loss of the coefficient, multiply \( \sqrt{\frac{1}{m-1}} \) after the denominator taking the square root. Thus, the confidence formula can be finally determined as \( \sqrt{\frac{m-1}{m-1}} \text{Cov}(X,Y) \). The confidence of each indicator is shown in Table 6(a) (normalized results). The Pearson correlation coefficient was used as the comparison result, as shown in Table 6(b).

The comparison results show that the new association rules algorithm has more accurate confidence. Different from the traditional confidence description form, the mean value of row vector and the mean value of column vector are added in this paper, and their values are equal to the normalized result of the mean value of each indicator. The mean value of row vectors represents the driving effect of the current indicator on other indicators, and the mean value of column vectors represents the consumption frequency of the current indicator in the five indicators. From this, the following main conclusions can be drawn from Table 6.
### Table 6(a). The Confidence Results of Five Indicators

| Indicators | Food     | Clothing | Education | Medical | Housing | Mean  |
|------------|----------|----------|-----------|---------|---------|-------|
| Food       | × 76.14% | 83.10%   | 85.89%    | 95.78%  | 19.46%  |
| Clothing   | 98.04%   | × 81.22% | 83.98%    | 93.20%  | 20.35%  |
| Education  | 100.00%  | 75.91%   | × 86.16%  | 95.68%  | 20.42%  |
| Medical    | 99.57%   | 75.62%   | 83.01%    | × 94.96%| 20.16%  |
| Housing    | 99.87%   | 75.47%   | 82.90%    | 85.40%  | × 19.62%|
| Mean       | 22.69%   | 17.30%   | 18.85%    | 19.49%  | 21.67%  |

### Table 6(b). Comparison Results of Pearson Correlation Coefficients

| Indicators | Food     | Clothing | Education | Medical | Housing |
|------------|----------|----------|-----------|---------|---------|
| Food       | × 97.93% | 99.89%   | 99.46%    | 99.75%  |
| Clothing   | 97.93%   | × 97.63% | 97.26%    | 97.06%  |
| Education  | 99.89%   | 97.63%   | × 99.78%  | 99.65%  |
| Medical    | 99.46%   | 97.26%   | 99.78%    | × 98.90%|
| Housing    | 99.75%   | 97.06%   | 99.65%    | 98.90%  |

### 3. Application of evaluation methods and data analysis

This paper uses this method to analyse the association rules among CPI, per capita consumption expenditure and per capita disposable income in rural areas. Association rules of each evaluation indicator can be obtained as shown in Table 7.

### Table 7. Association Rules for Each Evaluation Indicator

|                | Food CPI | Clothing CPI | Education CPI | Medical CPI | Housing CPI | Disposable Income |
|----------------|----------|--------------|---------------|-------------|-------------|-------------------|
| Food CPI       | -3.75%   | -3.18%       | -3.23%        | -3.82%      | -1.17%      | -2.47%            |
| Clothing CPI   | 3.43%    | 2.31%        | 3.19%         | 3.56%       | 0.78%       | -1.24%            |
| Education CPI  | -0.47%   | -0.48%       | -0.34%        | -0.26%      | -0.11%      | 1.66%             |
| Medical CPI    | 8.66%    | 6.73%        | 6.99%         | 8.41%       | -0.97%      | 3.94%             |
| Housing CPI    | -5.95%   | -5.38%       | -5.00%        | -5.49%      | -1.49%      | -1.21%            |
| Food           | ×        | 76.14%       | 83.10%        | 85.89%      | 95.78%      | 19.46%            |
| Clothing       | ×        | 81.22%       | 83.98%        | 93.20%      | 19.46%      | 81.22%            |
| Education      | 100.00%  | × 86.16%     | 95.68%        | 95.78%      | 81.22%      | 100.00%           |
| Medical        | 99.57%   | 75.62%       | × 94.96%      | 95.78%      | 81.22%      | 99.57%            |
| Housing        | 99.87%   | 75.47%       | 82.90%        | × 95.78%    | 81.22%      | 99.87%            |
| Disposable     | 22.57%   | 17.99%       | 18.67%        | 19.22%      | 21.54%      | 22.57%            |

### 4. Conclusions

In this paper, a new improved association rule algorithm is applied to rural development evaluation in China, and finally, the following suggestions were obtained:

1. Further increase per capita disposable income. It can raise the consumption level of each evaluation indicator to different degrees. It can be seen that the improvement of consumption level further reduces the CPI index of the key evaluation indicators, thus stabilizing the living standard of residents.

2. Appropriately reduce the price of medical care and clothing and promote the consumption level of residents in these two areas. It can be seen that this will reduce the fund transfer of these two indicators to other indicators, thus improving the richness of residents' material and spiritual life.
3) Maintain price stability for all evaluation indicators, especially housing, food and education. It can be seen that there is a positive correlation between the key evaluation indicators. Stabilizing prices can indirectly maintain residents' purchasing power towards other indicators.

5. Acknowledgments
This work was financially supported by “Cangzhou key research and development plan guidance project (Grant No. 172103003)” fund.

References
[1] Rakesh Agrawal, Tomasz Imieliński, Arun Swami. Mining association rules between sets of items in large databases[J]. ACM SIGMOD Record, 1993, 22(2).
[2] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, in 20th Int. Conf. Very Large Data Bases (Santiago, Chile), Morgan Kaufmann Publishers Inc., San Francisco, CA, 1994, 487–489.
[3] Mateen R. Shaikh, Paul D. McNicholas, M. Luiza Antonie, Thomas Brendan Murphy. Standardizing interestingness measures for association rules[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2018, 11(6).
[4] Jiawei Han, Jian Pei, Yiwen Yin. Mining frequent patterns without candidate generation[J]. ACM SIGMOD Record, 2000, 29(2).
[5] Longtao Liu, Jiaobao Wen, Zexun Zheng, Hansong Su. An improved approach for mining association rules in parallel using Spark Streaming[J]. Circuit Theory and Applications, 2021, 1.
[6] Seyed Mohsen Ghafari, Christos Tjortjis. A survey on association rules mining using heuristics[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(4).
[7] N. Marín, M.D. Ruiz, D. Sánchez. Fuzzy frameworks for mining data associations: fuzzy association rules and beyond[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2016, 6(2).
[8] Prasad S. Nishtala, Te-yuan Chyou, Fabian Held, David G. Le Couteur, Danijela Gnjidic. Association rules method and big data: Evaluating frequent medication combinations associated with fractures in older adults[J]. Pharmacoeconomics and Drug Safety, 2018, 27(10).
[9] Angelos Chatzimparmpas, Stamatia Bibi. Maintenance process modeling and dynamic estimations based on Bayesian networks and association rules[J]. Journal of Software: Evolution and Process, 2019, 31(9).
[10] James Xue, Stephen Jarvis. Mining association rules for admission control and service differentiation in e-commerce applications[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(3).
[11] Bapna Manish. World poverty: Sustainability is key to development goals[J]. Nature, 2012, 489(7416).
[12] Mingyue Liu, Xiaolong Feng, Sangui Wang and Huanguang Qiu. China’s poverty alleviation over the last 40 years: successes and challenges[J]. Agricultural and Resource Economics, 2020, 64(1).
[13] Jikun Huang, Scott Rozelle, Xinkai Zhu, Shiji Zhao, Yu Sheng. Agricultural and rural development in China during the past four decades: an introduction[J]. Australian Journal of Agricultural and Resource Economics, 2020, 64(1).
[14] NSBC (2019). China Rural Statistical Yearbook, China Bureau of Statistics. China Statistical Press, Beijing.
[15] International Fund for Agricultural Development. (2014). Rural poverty approaches, policies and strategies in China. Rome: IFAD.
[16] Walter Bossert, Satya R. Chakravartty, Conchita D’Ambrosio. Multidimensional Poverty and Material Deprivation with Discrete Data[J]. Review of Income and Wealth, 2013, 59(1).