OPEN ACCESS
EURASIA Journal of Mathematics Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2017 13(10):6555-6564
DOI: 10.12973/ejmste/78097

Educational Evaluation Based on Apriori-Gen Algorithm

Chen-Lei Mao
School of Management, Jiangxi University of Technology, Nanchang 330098, Jiangxi, CHINA
Song-Lin Zou
School of Management, Jiangxi University of Technology, Nanchang 330098, Jiangxi, CHINA
Jing-Hai Yin
School of Management, Jiangxi University of Technology, Nanchang 330098, Jiangxi, CHINA

Received 22 April 2017 ▪ Revised 5 August 2017 ▪ Accepted 3 September 2017

ABSTRACT
The issue of educational evaluation has long been a research hotspot. Using big data analysis method to conduct educational evaluation can improve the pertinence and effectiveness of education. Conventional Apriori algorithm has certain limitations in the application of educational evaluation. This paper introduces an improved Apriori-Gen algorithm and describes its application in evaluation of actual effectiveness of ideological and political course of colleges and universities. Through conducting correlation analysis of network questionnaire data, the study requirements of college students can be acquired, so as to improve the teaching effectiveness of ideological and political course. Results show that it is effective to apply the proposed study method in educational evaluation.

Keywords: big data, effectiveness evaluation, Apriori, ideological and political course

INTRODUCTION

Implementation of reasonable educational evaluation is the premise for education decision making. An effective education evaluation relies on a comprehensive and solid evaluation basis. Big data stresses on in-depth mining and analysis of multidimensional data so as to seek the implication relation and value behind data, which is beneficial for transforming educational evaluation from prediction based on small data to evidential decision based on comprehensive data. With the aid of big data technology, educational evaluation is no longer made to support the decisional requirement of education management departments or education institutions only, but for all groups and individuals that are concerned about education or taking parts in education. Through analyzing students' study requirements via big data, the pertinence and effectiveness of education can be improved.

The information found by data mining is normally meaningful knowledge that is impossible to be found by manual power. Data mining algorithms are many, including Apriori (Agrawal and Shafer, 1996; D'Angelo et al., 2016), K-means (Scitovski and Sabo, 2014), SVM (Support Virtual Machine) (Hu et al., 2015; Mu et al., 2017), EM (Expectation-Maximization) (Enders, 2003), Pagerank (Chen et al., 2007), Adaboost (Adaptive Boosting) (Hu, 2017a), KNN (K-Nearest Neighbor) (Hu, 2017b), Naive Bayes (Sitthi et al., 2016), etc.

Data are regarded to be associated when there is a certain regularity among them. The types of association are various, including simple association, chronicle association, causality association, quantitative association, etc. The purpose of association analysis is to find the correlation relation behind data. The association rule mining is to find meaningful and valuable association relation between item and set in database. In 1993, Agrawal et al.

Correspondence: Chen-Lei Mao, *School of Management, Jiangxi University of Technology, China.*
271053192@qq.com

Contribution of this paper to the literature

- This paper introduces an improved Apriori-Gen algorithm and describes its application in evaluation of actual effectiveness of ideological and political course of colleges and universities.
- The improved Apriori-Gen algorithm modified the bias during the teaching process and improved the teaching effectiveness of ideological and political education.

proposed for the first time the item-item association relation in mined database. Since then, many researchers conducted further studies on the association rule proposed by Agrawal et al. such as algorithm optimization, and introducing sampling and concurrent thought to improve algorithm efficiency.

The association rule mining proposed by Apriori contains two main parts: (1) to find all frequent itemsets in database according to a given minimum support; (2) to produce association rules. The key of the first part is to efficiently list all qualified frequent itemsets, which is also the most important issue in association rule mining technology. The improvement trend of association rule mining algorithm is to find all frequent itemsets that meet minimum support threshold.

To optimize Apriori's method, many research teams successively proposed various improvement thoughts. Holt et al. proposed IHP algorithm (Holt and Chung, 2002), in which the thought was to disperse the to-do-list into a hash table; Zaki et al. proposed Max Clique serial algorithm (Zaki, 1997), in which the thought was to utilize a clustering technique; Orlando et al. proposed DCP algorithm (Orlando et al., 2001), which can store and count candidate itemsets in a new way and integrate a more efficient pruning technique; Park et al. proposed DHP algorithm (Park et al., 1995), which can reduce the cost for generating candidate itemsets using the hash technology; Agarwal et al. proposed Tree Projection algorithm (Agarwal et al., 2001), in which ordered tree and mine frequent itemsets were constructed using database mapping technology; Savasere et al. proposed PARTITION algorithm (Savasere et al., 1995), which can cut database into random blocks, allowing each block individually to generate frequent itemsets; Toivonen et al. proposed Sampling algorithm (Toivonen, 1996), which can reduce the scale of frequent itemsets via a sampling technique. All these algorithms can improve data mining efficiency to a certain extent.

The bottleneck problem of itemsets generation may be encountered when using conventional Apriori algorithm, because there may generate too many candidate itemsets as well as massive amount of rule algorithms caused by repeatedly scanning the database (Song et al., 2006). How to select out interesting and valuable rules to be applied in practical situation has become a difficult issue. On the basis of analyzing the conventional algorithm, this paper proposes an improved Apriori algorithm, and elaborates the design thought, main problems and implementation method of the improved algorithm. Finally, the application of the improved Apriori association rule algorithm in ideological and political course is illustrated according to the actual educational environment of ideological and political course.

TECHNOLOGIES RELEVANT TO BIG DATA ANALYSIS

Big data mining is to mine valuable and potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, and random database, which is also a decision support process. Big data mining is mainly based on artificial intelligence, machine learning, pattern learning, and statistics. Common big data mining methods include classification, regression analysis, clustering, association rule, neutral network method, Web data mining, etc. These methods realize data mining from different perspectives. Association rule refers to the association and mutual relation between data items, which means that the generation of one data item can be used to deduce the generation of another item. The mining process of association rule mainly includes two stages: The first stage is to search all high-frequency itemsets from massive amounts of original data; the second stage is to generate association rule from these high-frequency itemsets.

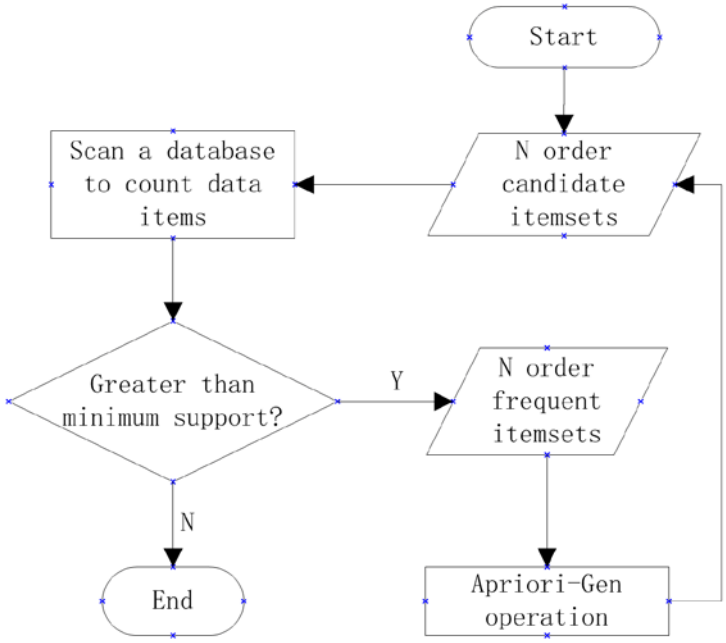

Figure 1. The flow chart of Apriori-Gen algorithm

Apriori algorithm is the most classic method for association analysis, in which the principle is as follow: If one itemset is frequent, then all its subsets must be frequent, i.e., if the current itemset is not frequent, then its superset will no longer be frequent.

The algorithm steps are described as below:

(1) At the beginning stage of the algorithm, determine the support degree of each item through one-pass scanning of the dataset. After finishing this step, all frequent 1 itemsets and set F1 can be obtained;

(2) Subsequently, this algorithm generates new candidate k itemset using frequent (k-1) itemset found by the last iteration;

(3) To count the support degree of candidate itemset, the algorithm needs to re-scan the database, and use subset function to determine all candiate k itemsets in C_k of each object t;

(4) After calculating the support degree count of candidate itemset, the algorithm will delete all candidate itemsets with support degree count less than minsup;

(5) Algorithm stops upon no generation of new frequent itemsets.

Concrete algorithm is shown as **Figure 1**.

METHOD

In the association analysis-based data mining algorithm, the three most important procedures are data collection, data preprocessing, and data analysis:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Serial nu	grade	major	Gender	Political c	NO.5 A	NO.5 B	NO.5 C	NO.5 D	NO.5 E	NO.6	NO.7	NO.8
2	1	2	1	2	3	1	1	1	1	1	1	1	1
3	2	3	1	1	3	1	1	1	1	1	1	2	1
4	3	3	2	2	2	0	1	0	0	0	2	2	2
5	4	3	1	2	2	1	1	1	1	1	2	2	2
6	5	3	2	2	3	1	1	1	1	1	1	2	1
7	6	3	2	2	3	1	1	1	1	1	1	2	1
8	7	3	2	2	1	1	1	1	1	1	1	2	1
9	8	3	1	1	3	1	1	1	1	1	1	1	3
10	9	3	1	2	3	1	1	0	1	1	2	2	3
11	10	3	1	2	3	0	1	1	1	1	1	1	1
12	11	3	2	2	3	0	1	1	1	1	2	2	1
13	12	3	1	2	2	1	1	1	1	1	2	2	1
14	13	3	1	1	3	1	1	0	1	1	1	2	1
15	14	3	1	1	3	1	1	1	1	1	1	2	3

Figure 2. Original questionnaire dataset

Collection of Data

College students' basic evaluations on the teaching effectiveness of ideological and political course were obtained by means of a questionnaire. The questionnaire included 47 questions, covering learner factors, teacher factors, and environment factors, etc. The questionnaire was implemented on website and the resulting data were exported in an Excel form as shown in **Figure 2**.

The first line is the name of each question, below which is the index value of item. To reduce the redundancy of storage, the content of each item is stored in another file.

Data Preprocessing

First, the questionnaire data in excel was subjected to processing treatment, wherein the question code was added in front of index value of each question, so that the index value of different questions can be distinguished from each other. After that, the questionnaire data was imported into RStudio using excel toolkit of R language. Data set is shown in **Figure 3**.

mydataEng ×

Filter

	number	question 1	question 2	question 3	question 4	question 5	question 6	question 7	question 8	quest
1	1	question 1-2	question 2-1	question 3-2	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
2	2	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
3	3	question 1-3	question 2-2	question 3-2	question 4-2	question 5-0	question 6-1	question 7-0	question 8-0	qu
4	4	question 1-3	question 2-1	question 3-2	question 4-2	question 5-1	question 6-1	question 7-1	question 8-1	qu
5	5	question 1-3	question 2-2	question 3-2	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
6	6	question 1-3	question 2-2	question 3-2	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
7	7	question 1-3	question 2-2	question 3-2	question 4-1	question 5-1	question 6-1	question 7-1	question 8-1	qu
8	8	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
9	9	question 1-3	question 2-1	question 3-2	question 4-3	question 5-1	question 6-1	question 7-0	question 8-1	qu
10	10	question 1-3	question 2-1	question 3-2	question 4-3	question 5-0	question 6-1	question 7-1	question 8-1	qu
11	11	question 1-3	question 2-2	question 3-2	question 4-3	question 5-0	question 6-1	question 7-1	question 8-1	qu
12	12	question 1-3	question 2-1	question 3-2	question 4-2	question 5-1	question 6-1	question 7-1	question 8-1	qu
13	13	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-0	question 8-1	qu
14	14	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
15	15	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
16	16	question 1-3	question 2-2	question 3-1	question 4-2	question 5-1	question 6-1	question 7-1	question 8-1	qu
17	17	question 1-3	question 2-2	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
18	18	question 1-3	question 2-2	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu
19	19	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-0	question 8-1	qu
20	20	question 1-3	question 2-1	question 3-1	question 4-3	question 5-1	question 6-1	question 7-1	question 8-1	qu

Figure 3. Data imported into RStudio

Before association analysis, the questionnaire data should first be converted into transaction data form. Therefore, the data was first converted into List form, then converted into transaction form. Key codes are shown below:

```
dataList <- split(data, f)
  dataList <- lapply(dataList, function(x){
    rst <- unlist(x)
    names(rst) <- NULL
    rst <- unique(rst)})
  transaction <- as(dataList, “transactions”)
```

Data Analysis

After being converted into transaction form, the association analysis of data was conducted. The key statement of apriori algorithm is as below:

```
rules = apriori(transaction, parameter = list(sup = 0.2, conf = 0.9))
```

where the minimum support degree was set to 0.2, the minimum confidence coefficient was set to 0.9, and the results are shown as follows:

```
> summary(rules)
set of 165077 rules
rule length distribution (lhs + rhs):sizes
2         3         4         5         6         7         8         9
75        2432      18618     53274     61087     26739     2847      5
Min.      1st Qu.   Median    Mean      3rd Qu.   Max.
2.000     5.000     6.000     5.602     6.000     9.000
```

We can see that there are over 160,000 qualified association rules, of which there are only 75 rules with a length of 2, and more than 2000 rules with a length of 3; rules with length of over 3 are too many, which will not be analyzed in this research. The algorithm parameters were modified, where the maximum number of association rules was set to 3, which means only association rules like A->B and A&B->C can be exported.

```
> myrules = apriori(transaction, parameter = list(maxlen = 3, sup = 0.2, conf
  = 0.9))
> myrules.sorted <- sort(myrules, by = “lift”)
> inspect(myrules.sorted)
```

The key codes are shown above. The analysis results were ranked according to their lift degrees and the following rules can be obtained:

```
      lhs                                rhs             support   confidence lift
[1]   {question 1-1,question 7-0}   => {question 8-0}  0.2136752 0.9174312  3.799627
[2]   {question 4-3,question 8-0}   => {question 1-1}  0.2179487 0.9714286  2.789132
[3]   {question 6-0,question 8-0}   => {question 1-1}  0.2222222 0.9629630  2.764826
[4]   {question 5-1,question 8-0}   => {question 1-1}  0.2222222 0.9629630  2.764826
[5]   {question 7-0,question 8-0}   => {question 1-1}  0.2136752 0.9345794  2.683332
[6]   {question 8-0}                => {question 1-1}  0.2222222 0.9203540  2.642489
[7]   {question 4-3,question 8-0}   => {question 7-0}  0.2158120 0.9619048  2.074523
[8]   {question 1-1,question 8-0}   => {question 7-0}  0.2136752 0.9615385  2.073733
[9]   {question 5-1,question 8-0}   => {question 7-0}  0.2200855 0.9537037  2.056836
[10]  {question 1-1,question 8-0}   => {question 6-0}  0.2222222 1.0000000  2.052632
[11]  {question 8-0}                => {question 7-0}  0.2286325 0.9469027  2.042168
[12]  {question 6-0,question 8-0}   => {question 7-0}  0.2179487 0.9444444  2.036866
[13]  {question 5-1,question 8-0}   => {question 6-0}  0.2286325 0.9907407  2.033626
[14]  {question 4-3,question 8-0}   => {question 6-0}  0.2222222 0.9904762  2.033083
[15]  {question 1-1,question 7-0}   => {question 6-0}  0.2286325 0.9816514  2.014969
```

The first 15 rules ranked in descending order of lift degree are given above, where the largest lift degree reaches 3.800.

RESULTS

First, the visualization analysis of all rules was carried out, where all 2507 effective rules were grouped and displayed in the form of a bubble diagram.

As shown in **Figure 4**, the x-coordinate represents the left operation of grouping rule, y-coordinate represents the right operation of grouping rule, the circle size represents support degree (the larger the circle, the larger the support degree), the color represents the lift degree of rule (the darker the color, the higher the lift degree). The relative frequencies of rules with support degree over 0.5 were recorded, as shown in **Figure 5**.

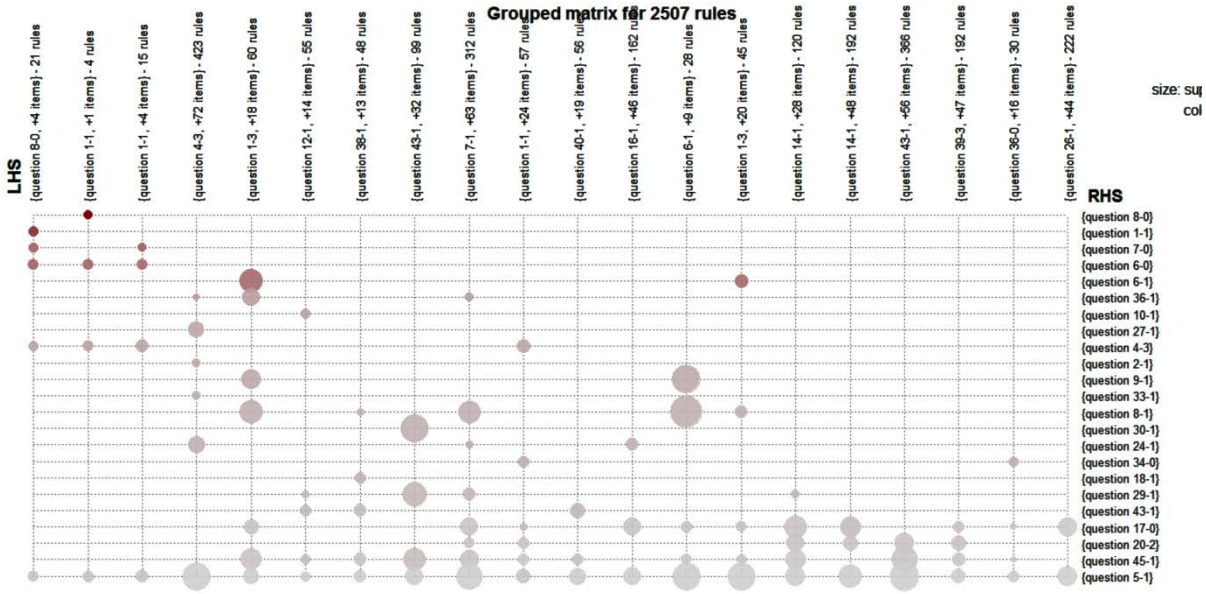

Figure 4. Bubble diagram of grouped rules

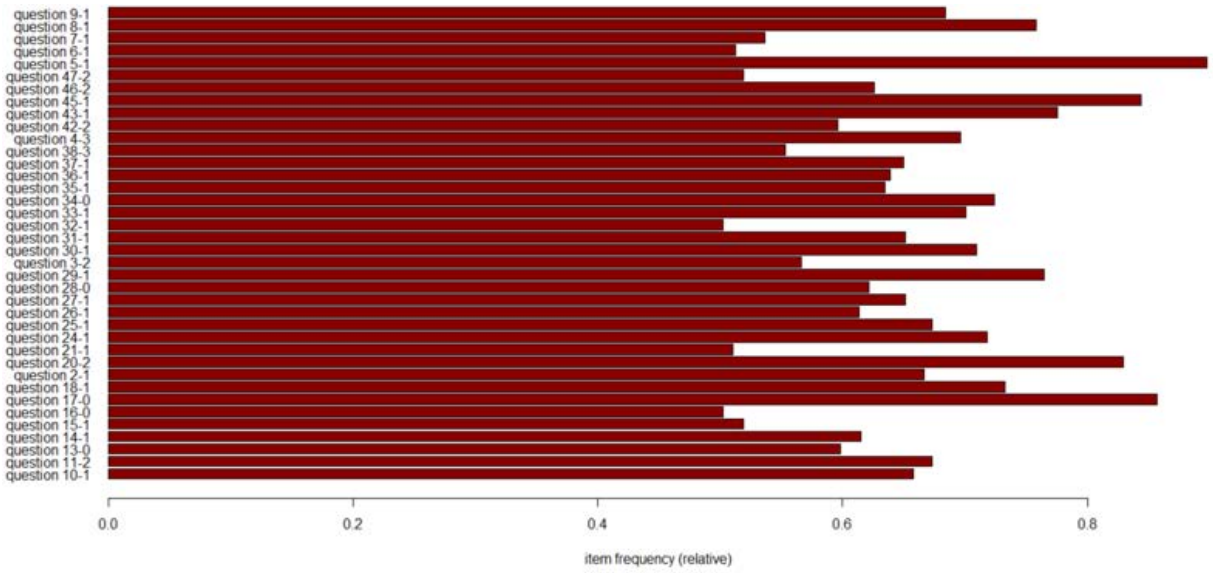

Figure 5. Relative frequencies of rules

```
> itemFrequencyPlot(transaction, support = 0.5, cex.names = 1, xlim = c(0,0.9),
  + type = 'relative', horiz = TRUE, col = 'dark red', las = 1)
```

We can conduct rule analysis for a single particular result. For example, the representative informant in question 1-1 is a college student, his/her association rules was specifically analyzed, and the results are shown below:

```
> rstRule <- Rhs_Selecet(myrules, "question 1-1")
> inspect(rstRule)
```

```
     lhs                              rhs             support   confidence lift
[1] {question 8-0}                 => {question 1-1} 0.2222222 0.9203540  2.642489
[2] {question 7-0,question 8-0}    => {question 1-1} 0.2136752 0.9345794  2.683332
[3] {question 6-0,question 8-0}    => {question 1-1} 0.2222222 0.9629630  2.764826
[4] {question 4-3,question 8-0}    => {question 1-1} 0.2179487 0.9714286  2.789132
[5] {question 5-1,question 8-0}    => {question 1-1} 0.2222222 0.9629630  2.764826
```

Figure 6 is a directed graph analysis of few key rules of a college freshmen student, where the arrow points to the direction of rule deduction. **Figure 7** is the scatter plot of rules, where the x-coordinate represents support degree, the y-coordinate represents confidence coefficient, the color of scattered point represents the lift degree (the darker the color, the higher the lift degree). Based on these rule analyses, the association relation between question and answer can be obtained combined with questionnaire results.

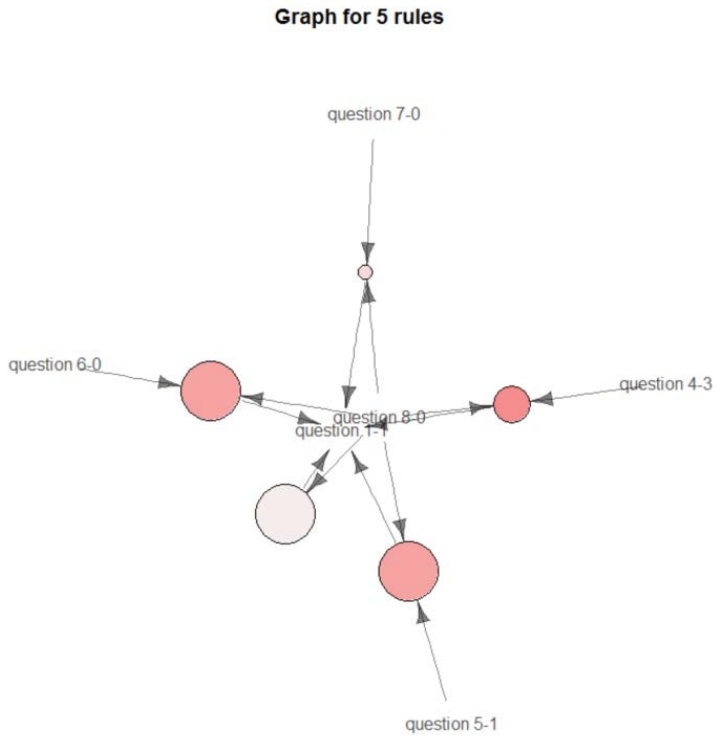

Figure 6. Digraph of specialized analysis of rules

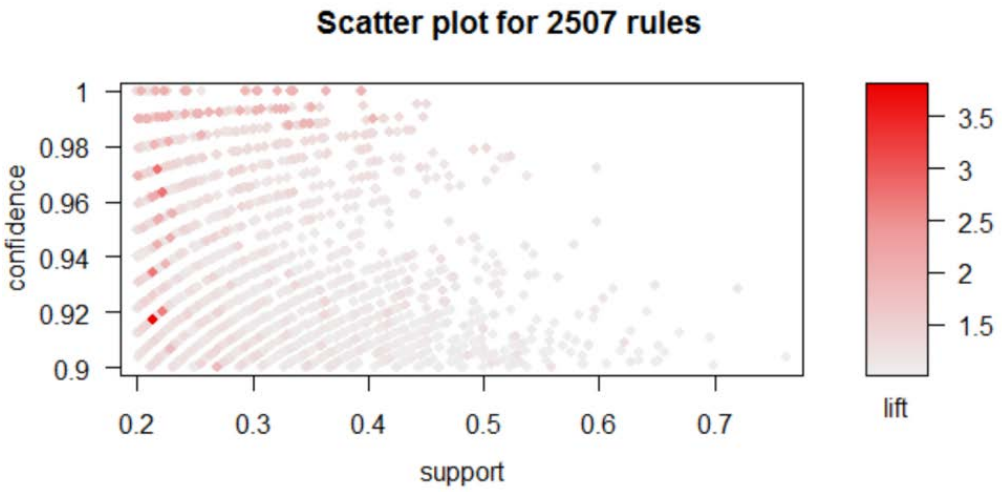

Figure 7. Scatter plot of rule analysis

DISCUSSION AND CONCLUSION

The evaluation of the effect of education has always been a focus of research. Varank et al. investigated the effectiveness of an online automated evaluation and feedback system that assessed students' word processing assignments (Varank et al., 2014). Öztürkler examined the current situation of the quality improvement in higher education institutions (Öztürkler, 2017).

Due to the limitation in evaluation conditions, traditional education evaluation normally collects only segmental evaluation information, and therefore may easily and passively ignore some evaluation points. During the implementation of educational evaluation, it will be over-reliant on subjective evaluation due to a lack of reliable evaluation basis. In contrast, big data-based educational evaluation does not rely on one-dimensional evaluation of a single evaluation object, but includes all contextual data related to education, not only using evaluation data, but also focusing on process data. The thought of seeking association via big data technique meets educational evaluation's true requirement for rich basis and valid evidence. The introduction of big data expands the content and function of educational evaluation, making it not only an evaluation, but also an important evidence for educational decision making. Peng used the big data processing technology on the online learning behavior analysis model (Peng, 2017).

However, due to the complexity of educational data, it is difficult to describe and analyze educational big data with general data analysis tools. With the development of higher education, it is ever more necessary to analyze and evaluate educational data so as to guide the formulation of educational policy and students' learning behavior. In this paper, 160000 association rules are extracted from educational data, which is too large for analyzing and evaluating data, so it must be condensed. By using the Apriori method, these association rules can be reduced to 2507 or even to 75. Although 75 seems simpler, this article uses 2507. After streamlining, these rules must be analyzed to group them to see which rules are more effective for evaluation, which rules may be redundant, which rules can be merged, and the set weights of the rules. The results of this study show that the research results are very satisfactory.

In this paper, an improved Apriori-Gen algorithm was applied for effective evaluation of ideological and political education, not only realizing overall evaluation, but also presenting individual situation effectiveness. The improved Apriori-Gen algorithm modified the bias during the teaching process and improved the teaching effectiveness of ideological and political education. By fully utilizing technical approaches, this new algorithm collected both students' study process and result data, and integrated various evaluation data (expert evaluation,

teacher evaluation, student self-evaluation, mutual evaluation), so as to realize multi-dimensional, comprehensive, in-depth evaluation of ideological and political education.

ACKNOWLEDGEMENT

This work is supported by the 2016 Ministry of Education of Humanities and Social Science project titled "Effectiveness Evaluation of College Ideological and Political Education Based on Big Data Technique (16JDSZ2052)".

REFERENCES

Agarwal, R., Aggarwal, C., & Prasad, V. V. V. (2001). A tree projection algorithm for generation of frequent itemsets. *Journal of Parallel and Distributed Computing, 61*(3), 350-371.

Agrawal, R., & Shafer, J. C. (1996). Parallel Mining of Association Rules. *IEEE Educational Activities Department*.

Chen, P., Xie, H., Maslov, S., et al. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics, 1*(1), 8-15.

D'Angelo, G., Rampone, S., & Palmieri, F. (2016). Developing a trust model for pervasive computing based on Apriori association rules learning and Bayesian classification. *Soft Computing*, 1-19.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8*(3), 322-37.

Holt, J. D., & Chung, S. M. (2002). Mining association rules using inverted hashing and pruning. *Information Processing Letters, 83*(4), 211-220.

Hu, J. F. (2017a). Automated detection of driver fatigue based on AdaBoost classifier with EEG signals. *Front. Comput. Neurosci*. doi:10.3389/fncom.2017.00072

Hu, J. F. (2017b). Comparison of Different Features and Classifiers for Driver Fatigue Detection Based on a Single EEG Channel. *Computational and Mathematical Methods in Medicine*. doi:10.1155/2017/5109530

Hu, J. F., Mu, Z. D., & Wang, P. (2015). Multi-feature authentication system based on event evoked electroencephalogram. *Journal of Medical Imaging and Health Informatics, 5*(4), 862-870.

Mu, Z. D., Hu, J. F., & Yin, J. H. (2017). Driving Fatigue Detecting Based on EEG Signals of Forehead Area. *International Journal of Pattern Recognition and Artificial Intelligence, 31*(5), 1750011.

Orlando, S., Palmerini, P., & Perego, R. (2001). Enhancing the Apriori Algorithm for Frequent Set Counting. *Data Warehousing and Knowledge Discovery*. Springer Berlin Heidelberg, 71-82.

Öztürkler, Z. (2017). Evaluation of Technology Strategies as Quality Strategy of Higher Education Institutions. *Eurasia Journal of Mathematics Science and Technology Education, 13*(7), 4021-4033. doi:10.12973/eurasia.2017.00770a

Park, J. S., Chen, M. S., & Yu, P. S. (1995). An effective hash-based algorithm for mining association rules. *In Proc. 1995 ACM SIGMOD Int. Conf. Management of Data*, 175-186.

Peng, W. (2017). Research on Online Learning Behavior Analysis Model in Big Data Environment. *Eurasia Journal of Mathematics Science and Technology Education, 13*(8), 5675-5684. doi:10.12973/eurasia.2017.01021a

Sarasere, A., Omiecinsky, E., & Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. *In 21st Int. Conf. On Very Large Databases, Zurich, Switzerland*, 105-112.

Scitovski, R., & Sabo, K. (2014). Analysis of the k -means algorithm in the case of data points occurring on the border of two or more clusters. *Knowledge-Based Systems, 57*(2), 1-7.

Sitthi, A., Nagai, M., Dailey, M., et al. (2016). Exploring Land Use and Land Cover of Geotagged Social-Sensing Images Using Naive Bayes Classifier. *Sustainability, 8*(9), 921.

Song, Q. B., Sheppred, M., Cartwright, M., & Mair, C. (2006). Software Defect Association Mining and Defet Correction Effort Prediction. *IEEE Transactions on Software Engineering*, 69-82.

Toivonen, H. (1996). Sampling large databases for association rules. *In Proc. 1996 Int. Conf. Very Large Databases, Bombay, India*, 134-135.

Varank, İ., Erkoç, M. F., Büyükimdat, M. K., Aktaş, M., & Yeni, S. (2014). Effectiveness of an Online Automated Evaluation and Feedback System in an Introductory Computer Literacy Course. *Eurasia Journal of Mathematics, Science & Technology Education, 10*(5), 395-404.

Zaki, M. J., Parthasarathy, S., & Li, W. (1997). A localized algorithm for parallel association mining. *IEEE*, 321-330.