COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

A Study on the Application of Data Mining Techniques in the Management of Sustainable Education for Employment

RESEARCH PAPER

FANG FANG

]u[ubiquity press

ABSTRACT

With the gradual advancement of education management towards data and informationisation, how to establish a perfect employment education management system has become an important element of current student work. Data mining technology can extract valuable implicit information from the database, and is widely used in the practical processing of massive data. Based on the analysis of the characteristics of employment education management in universities, the study first improved the K-means algorithm by adding splitting and aggregation operations to it, used the improved K-means algorithm to cluster and analyse the employment education data, and then combined it with the optimised Apriori algorithm to further mine the useful information in the employment education data. The experimental outcomes demonstrate that the error rate of the improved K-means algorithm is stable at around 10%, with a high accuracy rate and strong stability; combined with the optimised Apriori algorithm applied to the employment education management system, the accuracy rate is basically maintained at over 96%, and the scores of students' employment knowledge and employment practice are all over 90, indicating that the method can provide effective guidance for students' employment and give a guideline for the sustainable education management of employment.

CORRESPONDING AUTHOR:

Fang Fang

Electromechanic Engineering College, Zhejiang Tongji Vocational College of Science and Technology, Hangzhou, 311231, China

FanggFang2023@outlook.com

KEYWORDS:

Employment education; Data mining; K-means; Apriori; Sustainability

TO CITE THIS ARTICLE:

Fang, F. 2023. A Study on the Application of Data Mining Techniques in the Management of Sustainable Education for Employment. *Data Science Journal*, 22: 23, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2023-023

I. INTRODUCTION

Career education, which focuses on career development, job selection and workplace planning, is an important part of the educational work being carried out in universities (Salal et al. 2019). With the development of modernisation and information technology, universities are gradually implementing information management in education, i.e., building information systems for students and teachers, and employment education is no exception. In today's employment education managerial system, the information associated with it is huge and complex, which directly affects the upbringing of students' employability and the sustainable advancement of employment education (Trakunphutthirak et al. 2022). As a data processing tool, data mining technology can be applied to education management by selecting appropriate analysis tools to process the information in the database to obtain useful and valuable information, which has a broad development prospect. Data mining commonly used methods include clustering, association rules and regression analysis, etc. In practical use, the appropriate method must be selected according to the characteristics of the database (Abu et al. 2019). The K-means algorithm typically utilized in clustering and the classical Apriori algorithm in association rules both face problems such as low efficiency and need some improvement. Therefore, the research is based on improving the K-means algorithm and the Apriori algorithm, and applying them together in the employment education management system with a view to improving its data management capabilities. The first part of the article is a literature review on data mining technology in educational data management, including the improvement and application of K-means algorithm and Apriori algorithm (Jeong et al. 2018). The second part describes the improved K-means and Apriori algorithms in detail, and the third part is the verification of the application effect of data mining technology in employment education management, including the respective verification of the improved algorithms and the combined practical application effect verification. By verifying the application effect of data mining technology in employment education management, we hope to obtain more effective methods to further optimise management.

II. RELATED WORK

The key to sustainable education management in employment is the efficient processing of all relevant data and the mining of valuable information. In recent years, improvements to the K-means algorithm have received much attention from professionals and the research outcomes have been very fruitful. Lakshmi K's research team, to cope with the local optimal solution caused by the casual election of the incipient prime of the K-means algorithm, proposed to apply a population-based metaheuristic optimisation algorithm to the upgraded K-means algorithm, which is according to the intelligent behaviour of crows and is able to find the global top-notch key. Experiments in the benchmark dataset the outcomes demonstrate that the upgraded K-means algorithm has high accuracy (Lakshmi et al. 2018). Alguliyev, et al. (2020) developed a parallel clustering technique according to the K-means algorithm to increase the powerful computational power required for big data, which upgrades the clustering speed while maximising the preservation of the initial dataset characteristics and enables the clustering of the nearest centre of mass according to the obtained pivot of mass position. The effectiveness of this algorithm was verified in a comparison with the pre-improvement algorithm. Hossain, et al (2019) addressed the problem that the K-means algorithm has a high probability of grouping different items into the same group and designed a dynamic method for data clustering in which the K-means centre of mass is obtained by threshold calculation and the amount of clusters are formed with this value, thus enabling the data to be classified according to the comparison between the threshold and the Euclidean distance. The outcomes demonstrate that this method outperforms the pre-improvement method. Shrifan, et al. (2022) research team has optimised the K-means algorithm using Tukey's ordination in combination with a novel range calculation, to address the problem of large differences in data clustering accuracy due to different range calculations in classical K-means algorithm, which minimises the impact by eliminating outliers, and the outcomes demonstrate that the method significantly improves the convergence of the prime and increases the overall clustering accuracy by a value of 80.57%. Laxmi Lydia, et al. (2020) designed a new K-mean non-negative matrix decomposition method for the retrieval of valid information in big data, which incorporates a keyword extraction algorithm, and the outcomes demonstrate that the method reduces the error rate by 5%.

The team of Neysiani proposed to apply the butterfly optimisation algorithm, to association rule mining, to address the problem of low efficiency of existing data mining techniques, which used a parallel strategy of one CPU and three GPUs to run association rule mining, and used the CPU as a synchronizer (Neysiani et al. 2019). Wang and Zheng (2020) designed an upgraded Apriori for frequent itemset time series to address the problem of a large candidate set in the Apriori algorithm in data mining. Their outcomes demonstrated that it outperformed the traditional Apriori algorithm in terms of storage space based on the analysis of the time series relevance laws excavation process. Liu et al. (2021) for solving the problem that traditional data mining algorithms are difficult to mine large-scale data in a timely manner, they combined it with the frequent pattern growth algorithm in distributed parallel algorithms to achieve parallel mining of frequent itemsets and association rules, and the outcomes verified the efficient performance of the method (Liu et al.2021). Subha (2019) developed a distributed association rule mining algorithm for P-trees, which preserves transactional data through a special data structure P-trees. The experimental outcomes demonstrate that the method simplifies message exchange and database scanning and achieves lossless preservation of stored data. Sun (2019) applies data mining techniques to a university academic affairs management system and makes dynamic improvements based on the characteristics of the technique for mining potential information. The improved method improves the accuracy and constriction tempo of clustering and confirms the feasibility of the clustering method in computer network education management. In view of the sustainable development of college education, Wang and Soo-Jin (2021) used association rules to mine hidden data in student achievement information, and analyzed the influencing factors through classified decision tree Analysis of algorithms. The results show that the method can effectively optimize the teaching management system.

In summary, most researchers have improved the selection of initial centroids for the K-means algorithm, while optimising the Apriori algorithm accordingly. However, the clustering accuracy achieved is still low, and it is still difficult to meet the needs of educational data management. Therefore, K-means algorithm and Apriori algorithm are combined to further utilize data mining technology to process employment education data to achieve sustainable development of employment education management. Therefore, the improved Apriori algorithm is eventually combined with the SA-K-means method. Firstly, the SA-K-means algorithm is used to cluster the data to achieve pre-processing, and then the improved Apriori algorithm is used to mine the associated data in the data to better complete the employment education data classification and relationship mining.

III. DATA MINING-BASED SUSTAINABLE EDUCATION MANAGEMENT FOR EMPLOYMENT

A. CLUSTERING OF EMPLOYMENT EDUCATION MANAGEMENT DATA BASED ON THE K-MEANS ALGORITHM

Employment education in colleges usually involves employment education courses, the quality of employment education and students' professional performance, and the relationship between them is intricate and complex. At the same time, at this stage, universities are comprehensively strengthening employment and entrepreneurship education, focusing on cultivating students' comprehensive practical ability, and taking the improvement of comprehensive quality as the fundamental goal, tending to guide students to engage in self-employment and high-quality employment. The factors that need to be considered are gradually becoming complicated, and the huge amount of data and information generated makes it difficult for the employment education management system to handle efficiently (He 2021). In terms of the content of employment education alone, it mainly includes the establishment of career awareness, the development of employability, guidance on job search, guidance on self-employment education and so on. Therefore, according to the current educational development requirements, combined with expert consultation and theoretical analysis, and through the research of 32 famous schools in China, the study first established the employment education management system, and summarised the aspects of it that need to be processed by data, as demonstrated in Figure 1.

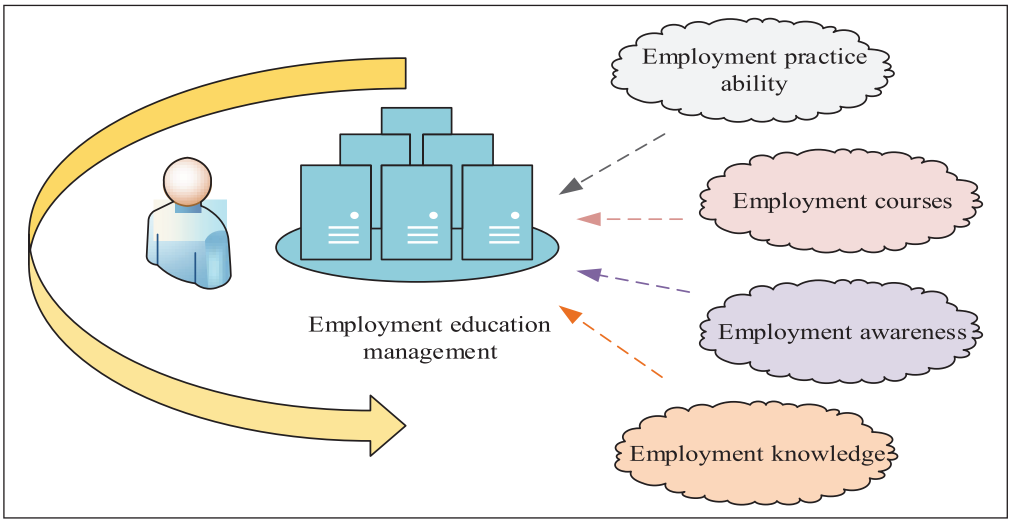

Figure 1 Main data processing of employment education management.

Before proceeding, the data is first clustered to be able to better mine the useful value information of employment education data (Thottathyl et al. 2020). Clustering analysis is a notable excavation of data excavation, and as one of the classical algorithms, the K-means is scalable, simple in principle, uncomplicated to exert, and has obvious behaviour advantages in data integration. The algorithm clusters all samples into the shortest distance clusters after determining k initial clustering centres, and achieves the optimum of the overall objective function under the action of continuous iteration (Xia 2021). However, the K-means algorithm usually requires human specification in determining the value of k, is highly dependent on the initial clustering centres, and has a high frequency of local optima. Therefore, the study incorporates splitting and aggregation operations into the K-means to form the SA-K-means algorithm to further improve the K-means method and make the data clustering outcomes more representative. the primordial framework of the K-means is demonstrated in Figure 2.

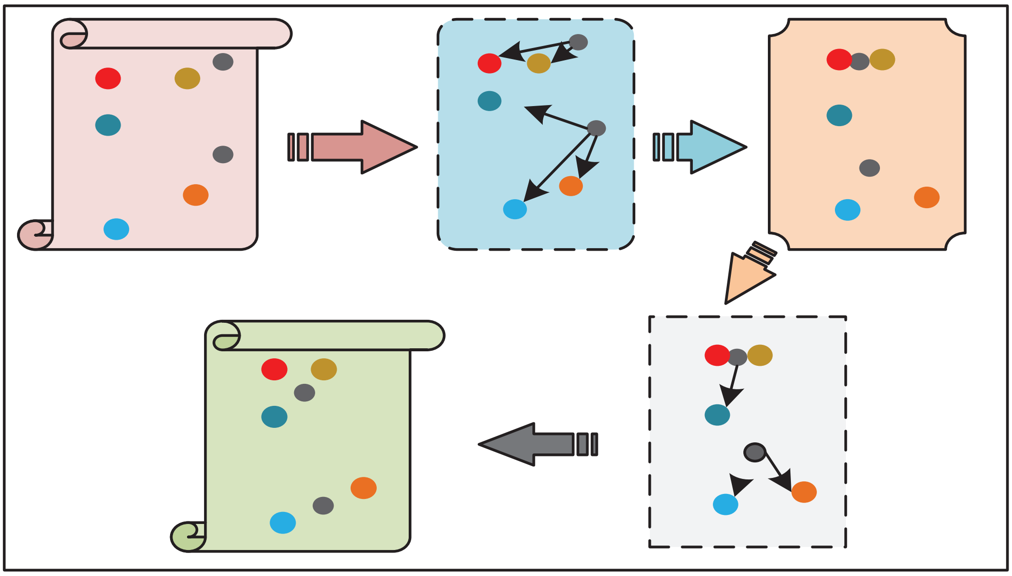

Figure 2 Basic principle of K-means algorithm.

The K-means for cluster analysis is an iterative process in which k sample data points are first selected in a random way to form an initial cluster centre in a certain data set (Bağdatli et al. 2021). The gap between the incipient cluster centre and the sample data points is the basis for the classification of the classes, i.e., the sample data points are classified into the nearest cluster centre according to the proximity principle. The new cluster centres are obtained by taking the mean of the attribute values of all data points in each class, and the amount of cluster centres are still k at this point. The final judgement is made on whether the clustering evaluation criterion function has reached the optimum, and if it has, then the class division continues, and if not, then it is iterated again. The objective criterion function is calculated as demonstrated in equation (1).

$$E = \sum_{j=1}^{c} \sum_{k=1}^{n_j} \left\| x_k - m_j \right\|^2 \tag{1}$$

In equation (1), E is the totality of the mean squared differences calculated from the attribute values of the data points, m_j is the cluster centre of the j cluster, and x_k is the individual data points in the sample data. For the classification criteria of the data samples, the similarity is followed and the Euclidean distance is used to determine the similarity, as demonstrated in equation (2).

$$d(x_i, x_j) = \sqrt{\sum_{k-1}^{n} (x_{ik} - x_{jk})^2} \tag{2}$$

In equation (2), x_i, x_j are samples contained in the dataset, and x_{ik} and x_{jk} are samples contained in the k cluster.

With the continuous execution of clustering, an optimal set of divisions is obtained. Not only does it maintain a maximum degree of independence between clusters, but a high degree of compactness is also maintained within individual clusters. Cluster analysis is judged by equation (3).

$$\begin{cases} J = \sum\limits_{i=1}^{k} \sum\limits_{\substack{j=1 \\ x_j \in C_i}}^{n} dis(x_j, c_i)^2 \\ c_i = \dfrac{1}{N_i} \sum\limits_{\substack{j=1 \\ x_j \in C_i}}^{n} x_j \end{cases} \tag{3}$$

In equation (3), c_i is the average of data, c_j is the data samples contained in the class C_i, the Euclidean distance between x_j and c_i is $dis(x_j, c_i)$ and N_i is the amount of data contained in the first i cluster. The vector of cluster centres needs to be corrected after the first clustering is completed, as demonstrated in equation (4).

$$z_j = \frac{\sum_{x \in s_j} x}{N_j}, j = 1,2,\ldots,k \tag{4}$$

In equation (4), z_j is cluster centre, N_j is the amount of samples contained in each of the different groups, and S_j is the group. The mean distance between the cluster centres and the samples is calculated by equation (5).

$$D_j = \frac{\sum_{x \in s} \left\| x - z_j \right\|}{N_j}, j = 1,2,\ldots,k \tag{5}$$

In equation (5), D_j is the average distance found. This gives the total average distance, see equation (6).

$$\overline{D} = \frac{1}{N} \sum_{i=1}^{k} \sum_{x \in s_j} \left\| x - z_j \right\| \tag{6}$$

In equation (6), $\overline{D}$ is the total mean distance. At this point a split or merge operation is added, which focuses on the outcomes of the previous clustering. The purpose of the splitting process is to increase the number of clustering centres as much as possible, while keeping the original clustering centres intact. A merge step is also added to deal with the problem of too close distances between data samples of different categories. The standard deviation between centroids and data samples is a necessary step for the added splitting operation, as demonstrated in equation (7).

$$\sigma_j = \sqrt{\frac{\sum_{x \in s_j} (x - z_j)^2}{N_j}} \tag{7}$$

In equation (7), σ is the standard deviation. For each grouping, there is a corresponding standard deviation, at which point the maximum of these is selected. When this maximum is bigger than the maximum of the standard deviation of the samples in the kind, the amount of samples in the kind exceeds the maximum value and the average distance is greater than the total average distance, or the number of clusters is less than one half of the required number, the classification operation is performed and the two sets of clustering centres are obtained, see equation (8).

$$\begin{cases} z_j^{+} = z_j - \rho\sigma_{\max}, 0 < \rho < 1 \\ z_j^{-} = \rho\sigma_{\max} + z_j \end{cases} \tag{8}$$

In equation (8), z_j^+ and z_j^- are the cluster centres obtained after splitting, and σ_{max} is the maximum standard deviation value. When performing the merge operation, the comparison of the cluster centres between the two groups is performed by equation (9).

$$D_{ij} = \left\| z_i - z_j \right\|, i = 1,2,3,\dots,k-1, j = i+1,\dots,k \tag{9}$$

When the minimum mean distance is smaller than the minimum value of the distance from the cluster centre, a merge operation is performed to obtain a new cluster centre as demonstrated in (10).

$$z_i^* = \frac{N_j z_j + N_i z_i}{N_j + N_i} \tag{10}$$

In equation (10), z_i^* is the new clustering centre. When the merging operation is completed, the number of corresponding clusters is subtracted by one, and finally the corresponding centroid vectors and groupings are output once the stop iteration condition is satisfied.

B. ANALYSIS OF EMPLOYMENT EDUCATION MANAGEMENT INFORMATION BASED ON ASSOCIATION RULES

After clustering the data information in employment education using the upgraded K-means method, the correlation between the information is further identified through association rules to facilitate educational information analysis, provide more scientific and effective guidance for employment education, and promote sustainable education development. The purpose of relevance laws is to extract all the strong correlation laws that exist in the target transaction database, i.e., the support of the association rules mined must meet the necessary conditions of greater than, or equal to, the least confidence, and the confidence level must be greater than, or equal to, the minimum confidence level (Safara et al. 2020). Finding frequent itemsets and computing deep correlation laws are the two cardinal procedures of relevance laws excavation, where the efficiency of correlation laws excavation is closely related to the efficiency of mining frequent itemsets (Zhan et al. 2019). The Apriori algorithm, as a classical method in relevant laws excavation, is primitive and uncomplicated to employ, and is widely used in transactional databases. The Apriori algorithm first sets the least support threshold and least confidence threshold based on the strength of the relevance rule, and then reads all the transaction data, in which all items are candidate 1 itemset C1. The support of all C1 items is then obtained from the total number of transactions and compared with the least support threshold one by one (Gupta et al. 2020). The one that is smaller than the minimum support threshold is removed and the one that is greater than or equal to it is kept as the frequent 1 itemset L1 and the candidate 2 itemset C2 is obtained by linking L1 with itself. A second scan of the database is then performed and the support of the C2 items is calculated and L3 is obtained and C3 is generated by following the steps after the first scan. The flow of Apriori algorithm prosperous itemset mining is demonstrated in Figure 3.

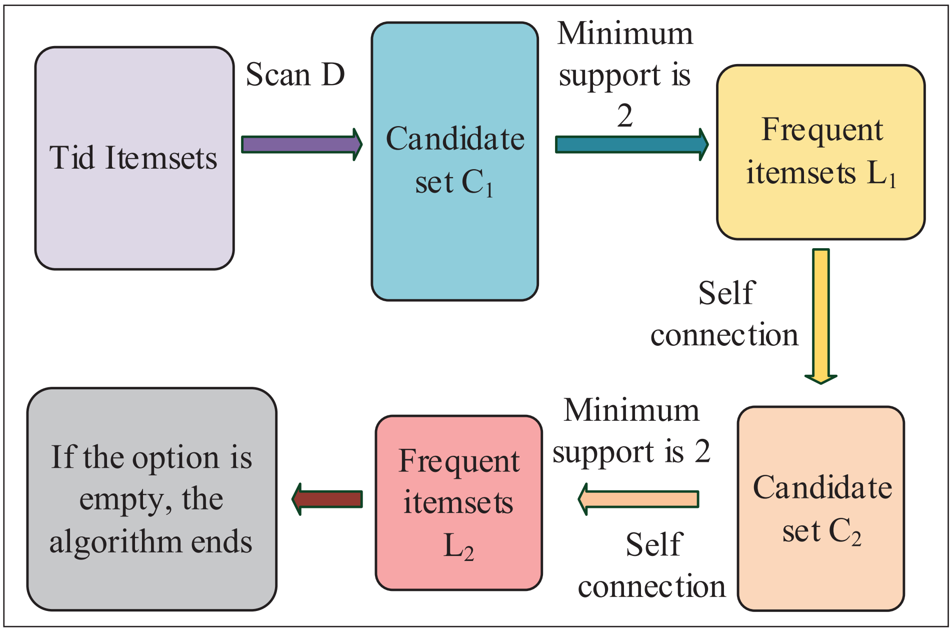

Figure 3 Mining process of complex itemsets in Apriori algorithm.

The calculation of the support of individual items is one of the key steps in the Apriori algorithm. The support actually represents the frequency of the set of items in a transaction and is calculated as demonstrated in equation (11).

$$\sup port(x) = \frac{number(x \subseteq T)}{|D|} \tag{11}$$

In equation (11), *D* is the transaction-specific database, *T* is the specific transaction, and Sup*port* is the support level and represents the subset of the outcoming itemset. After eliminating the unqualified items from the itemset by the support calculation, the frequent itemsets are merged to obtain the new set of options, as demonstrated in equation (12).

$$C_k = L_{(k-1)n} \cup L_{(k-1)m} \tag{12}$$

In Equation (12), C_k is the set of candidate items, *L* is the set of frequent items, and *n* and *m* represent the set of them. The new set of frequent items is then formed by combining the two sets and traversing them as demonstrated in equation (13).

$$C_j = C_k(j) \tag{13}$$

In equation (13), C_j is a subset of the outcoming new set of options and *j* is the *j* the subset of C_j. The size of the number of subsets is obtained by the calculation procedure demonstrated in equation (14).

$$count = \begin{cases} count, C_j \not\subset I_i \\ count+1, C_j \subset I_i (j = 0,1,2,\ldots)(i = 1,2,3,\ldots N) \end{cases} \tag{14}$$

In equation (14), C_j is a subset of C_k, *N* is the order, and *i* is the itemset from *i*. *count* Starting from 0, *k* is the maximum value of *j*. The confidence level is calculated as demonstrated in equation (15).

$$con(x \Rightarrow y) = \frac{\sup port(x \Rightarrow y)}{\sup port(x)} = \frac{P(xy)}{P(x)} = P(y|x) \tag{15}$$

In equation (15), *x* and *y* are both itemsets and are not equal, $con(x \Rightarrow y)$ is the odds of the make an appearance of the itemset *y* in the itemset *x* and $\sup port(x \Rightarrow y)$ is the probability of the data set containing both *x* and *y*, i.e., The probability of an itemset *P*(*xy*) *x* in the database is *P*(*x*) and the odds of itemset *x* containing itemset *y* is $P(y|x)$. It can be found that the Apriori algorithm still generates more non-essential candidate sets when the size of the candidate set is increasing (Da et al. 2019). Therefore, a parallel algorithm based on Matrix and Weight (MW-Apriori) is proposed to improve it. The algorithm introduces parallel computing and chunking of Boolean matrices according to the MapReduce framework, thus improving the productivity of the algorithm. The MW-Apriori algorithm also dwindles the amount of candidate sets by eliminating items that do not satisfy the conditions before performing join operations on the data, thus compressing the database.

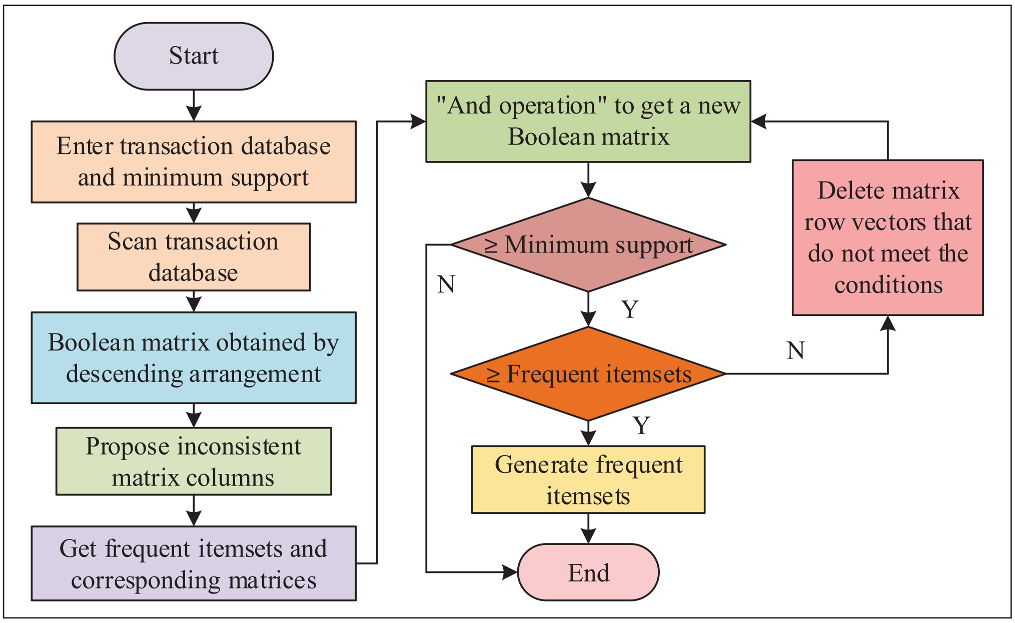
Start
Enter transaction database and minimum support
Scan transaction database
Boolean matrix obtained by descending arrangement
Propose inconsistent matrix columns
Get frequent itemsets and corresponding matrices
"And operation" to get a new Boolean matrix
≥ Minimum support
N
Y
≥ Frequent itemsets
N
Y
Delete matrix row vectors that do not meet the conditions
Generate frequent itemsets
End

Figure 4 The specific mining process of MW Apriori algorithm.

From Figure 4, the MW-Apriori algorithm stores database transactions based on the construction of a Boolean matrix, and uses frequency statistics to remove unqualified itemsets in advance, thereby constructing a new matrix. The matrix is then further compressed by finding different transactions with the same itemset and performing a merge weighting. The remaining matrix elements are then 'summed' two by two to obtain a matrix that meets the requirements, and finally the intersection of all the outcomes is used to obtain the association rules between the different data sets.

IV. ANALYSIS OF THE EFFECTIVENESS OF DATA MINING IN THE MANAGEMENT OF SUSTAINABLE EDUCATION FOR EMPLOYMENT

The research mainly applied the clustering algorithm and association rules in data mining for the processing of data related to employment education management. The SA-K-means algorithm was from improving the K-means for data clustering, and the upgraded Apriori method was to unearth the hidden relationships between data to provide guidance on employment education. The upgraded SA-K-means algorithm was therefore first analysed for performance, and the standard data, Iris, was selected to test it and compare it with the K-means method before the improvement. Iris dataset is a commonly used clustering experiment dataset, also known as iris flower dataset, which is a class of multivariate analysis dataset. The dataset contains 150 data samples, divided into three categories, with 50 data in each category, and each data contains four attributes. Fifty experiments were conducted on the upgraded K-means method before and after the experiment, and the amount of misclassified individuals and clustering outcomes were recorded. The clustering outcomes of the before and after algorithm are demonstrated in Figure 5. There are two types of clustering outcomes for the classical K-means algorithm, containing 37 test outcomes in Figure 5(a) and 13 test outcomes in Figure 5(b), and only one type of 50 test outcomes for the upgraded SA-K-means, which is the outcome in Figure 5(a). The K-means method is less stable, with two clustering outcomes, while the SA-K-means algorithm is relatively stable.

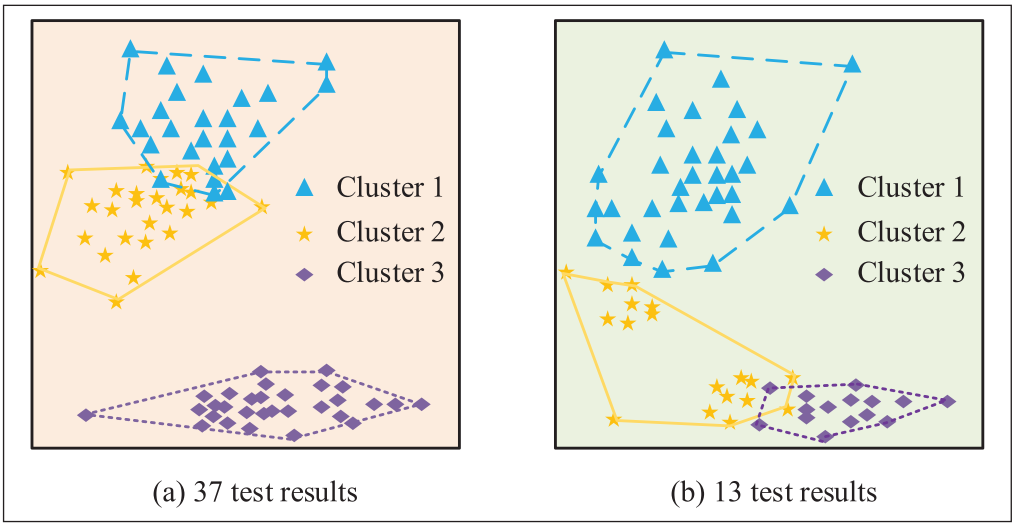

Figure 5 Clustering outcomes of K-means algorithm before and after improvement.

The outcomes of the wrong score rate of the K-means before and after the improvement are demonstrated in Figure 6. Figure 6 demonstrates the outcomes of the clustering error rate for the K-means method and the SA-K-means method. The horizontal coordinates are the batches of experiments, while the vertical coordinates represent the error rate. The error rate of the K-means ranges from 10% to 60 as the amount of experimental batches increases, with the error rate exceeding 50% in 13 instances and fluctuating widely. In contrast, the error rate of the optimised SA-K-means method was stable at around 10%, and did not change with the change of experimental batches, which was more steady and more pinpoint than the K-means. The effectiveness of the upgraded MW-Apriori algorithm was then verified.

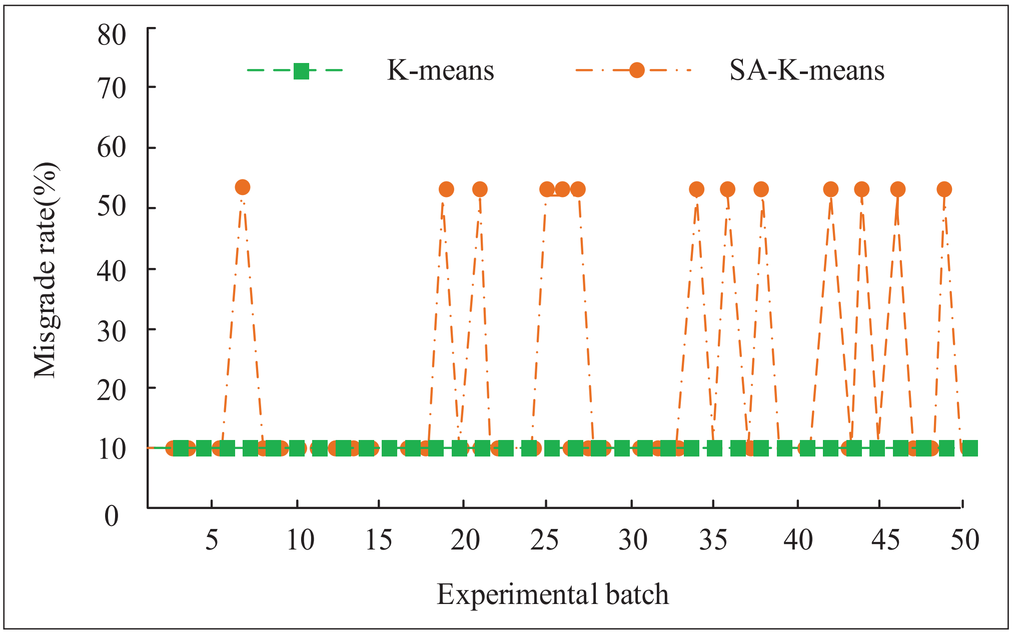

Figure 6 Outcomes of misclassification rate of K-means algorithm before and after improvement.

To elevate the credibility of the comparison, the upgraded method was compared with the classical Apriori method and the CM-Apriori means ground on the clustering matrix. All three methods were guaranteed to be tested and compared in the same experimental environment, as demonstrated in Table 1. The Mushroom dataset (Mushroom) was selected for the study to checkout the behaviour of the three means. The dataset contains 8,124 transactions, the amount of itemsets is 119, and the maximum length of the transactions is 23. The experiment consists of two parts; the first part is a comparison of the running times of the three methods in the dataset Mushroom for different minimum support degrees. To avoid memory overflow due to too small support threshold settings, the minimum support levels set in the study were 25%, 30%, 35%, 40%, 45%, 50%, and 55% respectively. The second part demonstrates the runtime discrepancy of the three methods for different amount of transactions in the dataset Mushroom with a support threshold of 30% and an increasing number of transactions in the dataset ranging from 2,500 to 5,500.

Table 1 Experimental software and hardware configuration of three methods.

SOFTWARE AND HARDWARE ENVIRONMENT CONFIGURATION	CONCRETE CONTENT
Operating system	Windows 10
Development platform	IntelliJ IDEA
Internal storage	4GB
Graphical tools	Matlab 2017b

The runtime outcomes for the three methods in the Mushroom dataset are demonstrated in Figure 7. Figure 7(a) demonstrates the runtime discrepancy of the three selected methods with different minimum support degrees, and Figure 7(b) demonstrates the runtime discrepancy of the three means with changing amount of transactions. From Figure 7(a), the runtime of the classical Apriori method fluctuates from 5s to 30s as the support level changes, with the longest running time of about 28s when the support level is 25% and the shortest running time occurring at 55% support level, which is roughly 8s. The longest and shortest running times of the CM-Apriori algorithm are 20s and 5s respectively, while the MW- Apriori algorithm has a maximum run time of 5s and a minimum time within 1s. The maximum runtime of the Apriori method and the CM-Apriori method are 30s and 22s respectively, while the maximum running time of the MW-Apriori algorithm is 10s, and compared to the first two methods, the upgraded MW-Apriori method is shorter and more efficient. Compared to the first two means, the upgraded MW-Apriori method has shorter time and higher running efficiency, with greater improvement in both time and space efficiency.

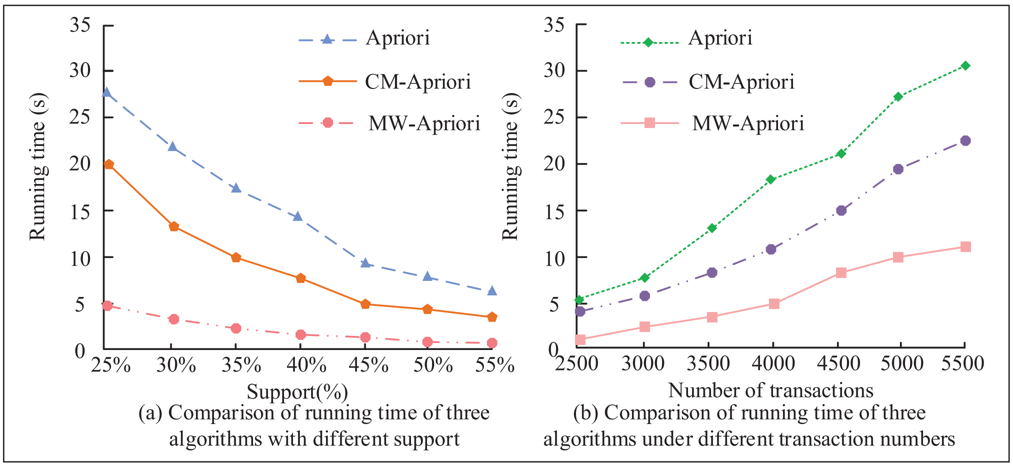

Figure 7 Running time outcomes of three methods in Mushroom dataset.

Finally, the SA-K-means method and MW-Apriori method were combined, named SAK-MWA method, and jointly applied to the employment education management, and the data mining effect of the combined SAK-MWA algorithm on the employment education-related data was first tested. The selected employment education management system was established by a well-known university in China itself, and involved various data of employment education of the university, including employment education contents, employment courses, etc. The mining outcomes of the three methods in this database are demonstrated in Figure 8. Figure 8 demonstrates the data mining accuracy outcomes of K-means method, Apriori method and the research combined SAK-MWA method in the employment education management system, the horizontal coordinates in Figure 8 are the database numbers consisting of randomly selected relevant data in this employment education management system. In Figure 8, the sort precision of the K-means method has large fluctuations and relatively poor stability, concentrated around 92%, while the accuracy of the Apriori method is relatively low, mostly at 88% and below. The accuracy of the combined SAK-MWA algorithm was maintained at 96% and above, with less fluctuation and better stability.

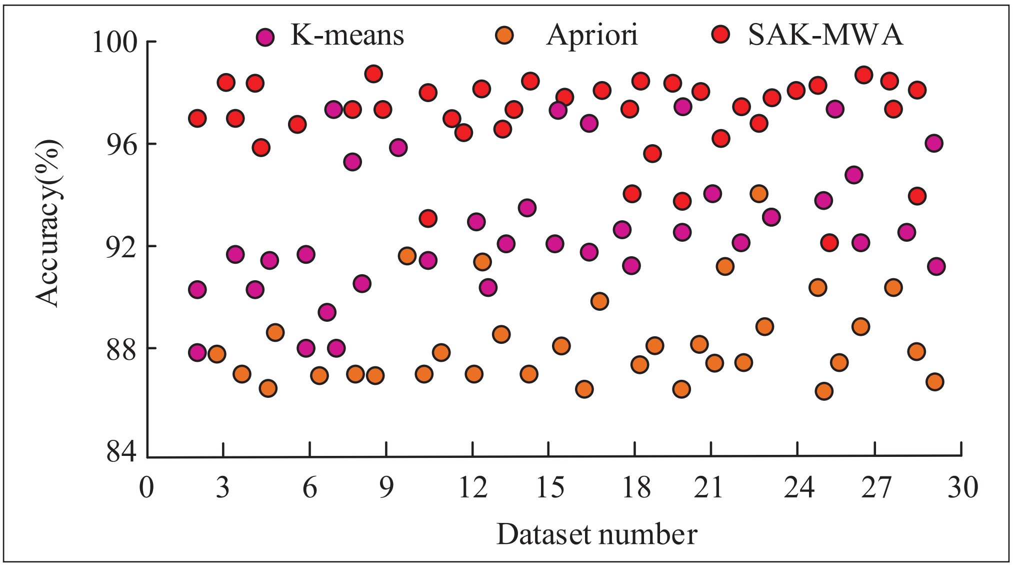

Figure 8 Comparison of data mining outcomes of three algorithms in employment education management system.

Finally, the SAK-MWA algorithm was applied to the employment education management system to evaluate the four aspects of employment knowledge mastery, employment practice ability, employment awareness, employment courses, and compared with the outcomes before use, as demonstrated in Figure 9. The horizontal coordinates in Figure 9 represent the four main aspects of employment education and the vertical coordinates are the ratings. From Figure 9, before the application of the SAK-MWA algorithm to the employment education management system, students' employment knowledge mastery, practical skills, employment awareness and employment course-related ratings were all at a low level, with only the employment course rating being close to 85. After the SAK-MWA algorithm was applied to the employment education management system, however, students' employability and knowledge mastery were significantly improved, with ratings above 90 in all four categories, including a score of 95 or more in employability practice, indicating that the method is conducive to a more efficient management of the employment education management system and provides effective assistance to students, which can promote the sustainable development of employment education.

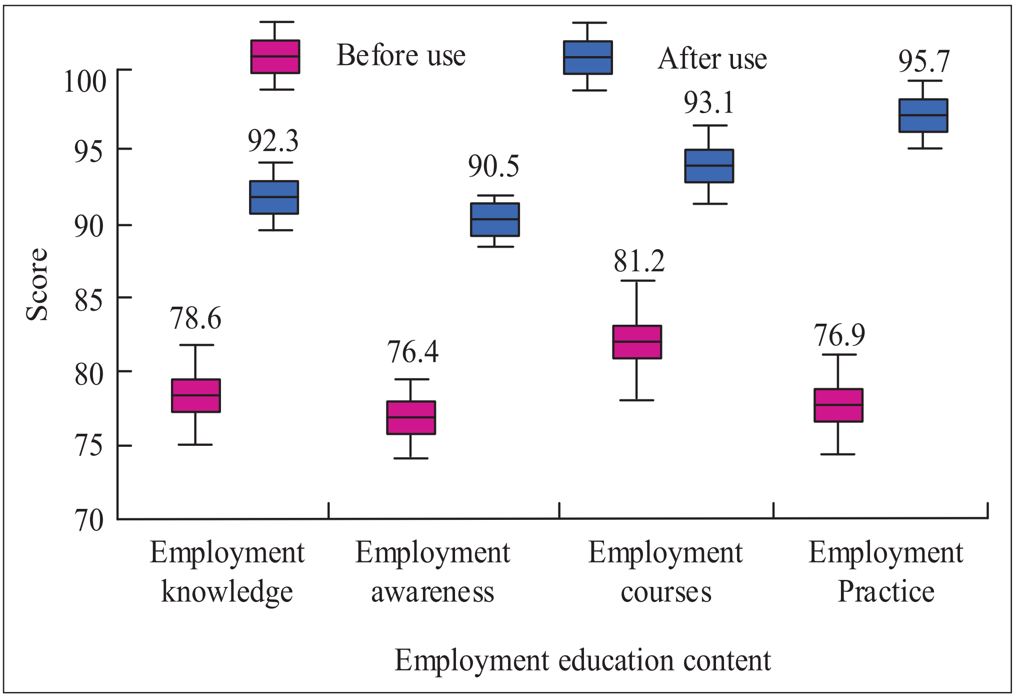

Figure 9 Comparison of the effectiveness of employment education management before and after using SAK-MWA algorithm.

V. CONCLUSION

Employment education is an indispensable element for universities to achieve quality development and is matter to the sustainable development of education. The study uses data excavation techniques to process information related to employment education by analysing the peculiarity of employment education management system. Firstly, the K-means is upgraded for clustering analysis of employment education data, and secondly, the Apriori algorithm is upgraded and the two are combined and applied together in the employment education management system. The outcomes demonstrate that the improved K-means algorithm has high stability with only one kind of clusters obtained in 50 tests, and its error score rate is kept near 10% with high accuracy; the optimised Apriori algorithm has a shortest running time of no more than 1s with different minimum support, and a longest running time of 10s when dealing with different number of transactions, which has a high running efficiency; the two improved algorithms were applied to employment education management, the students' employment practice ability was up to 95 points or more, and the employment knowledge acquisition, employment awareness and employment course evaluation were all above 90 points, indicating that the method has improved the efficiency of employment education management. However, the study did not optimise the self-linking and pruning steps of the Apriori algorithm when improving it, so further exploration in this area is needed.

FUNDING INFORMATION

The research is supported by Zhejiang China Vocational Education Research Project in 2022, Analysis of the Current Situation of the Innovation and Entrepreneurship Competition for Higher Vocational Students and Research on the Construction of Teaching Mode (No. ZJCV2022C15).

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Fang Fang orcid.org/0009-0009-3749-209X
Electromechanic Engineering College, Zhejiang Tongji Vocational College of Science and Technology, Hangzhou, 311231, China

REFERENCES

Abu Saa, A, Al-Emran, M and **Shaalan, K.** 2019. Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24(4): 567–598. DOI: https://doi.org/10.1007/s10758-019-09408-7

Alguliyev, RM, Aliguliyev, RM and **Sukhostat, LV.** 2020. Efficient algorithm for big data clustering on single machine. *CAAI Transactions on Intelligence Technology*, 5(1): 9–14. DOI: https://doi.org/10.1049/trit.2019.0048

Bağdatli, MEC and **Dokuz, AŞ.** 2021. Modeling discretionary lane-changing decisions using an improved fuzzy cognitive map with association rule mining. *Transportation letters*, 13(8): 623–633. DOI: https://doi.org/10.1080/19427867.2021.1919469

Da Costa, MB, Dos Santos, LMAL and **Schaefer, JL.** 2019. Industry 4.0 technologies basic network identification. *Scientometrics*, 121(2): 977–994. DOI: https://doi.org/10.1007/s11192-019-03216-7

Gupta, MK and **Chandra, PA.** 2020. Comprehensive survey of data mining. *International Journal of Information Technology*, 12(4): 1243–1257. DOI: https://doi.org/10.1007/s41870-020-00427-7

He, Q. 2021. Research on promoting the employment of college graduates with ideological and political education in the New Situation. *Advances in Educational Technology and Psychology*, 5(3): 141–145. DOI: 10.23977/aetp.2021.53020

Hossain, MZ, et al. 2019. A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(2): 521–526. DOI: https://doi.org/10.11591/ijeecs.v13.i2.pp521-526

Jeong, YJ, Lee, J, Moon, J, Shin, JH and **Lu, WD.** 2018. K-means data clustering with memristor networks. *Nano letters*, 18(7): 4447–4453. DOI: https://doi.org/10.1021/acs.nanolett.8b01526

Lakshmi, K, Visalakshi, NK and **Shanthi, S.** 2018. Data clustering using k-means based on crow search algorithm. *Sādhanā*, 43(11): 1–12. DOI: https://doi.org/10.1007/s12046-018-0962-3

Laxmi Lydia, E, et al. 2020. Charismatic document clustering through novel K-Means non-negative matrix factorization (KNMF) algorithm using key phrase extraction. *International Journal of Parallel Programming*, 48(3): 496–514. DOI: https://doi.org/10.1007/s10766-018-0591-9

Liu, L, et al. 2021. An improved approach for mining association rules in parallel using Spark Streaming. *International Journal of Circuit Theory and Applications*, 49(4): 1028–1039. DOI: https://doi.org/10.1002/cta.2935

Neysiani, BS, et al. 2019. Improve performance of association rule-based collaborative filtering recommendation systems using genetic algorithm. *International Journal of Information Technology and Computer Science*, 11(2): 48–55. DOI: 10.5815/ijitcs.2019.02.06

Safara, F, Souri, A and **Serrizadeh, M.** 2020. Improved intrusion detection method for communication networks using association rule mining and artificial neural networks. *IET Communications*, 14(7): 1192–1197. DOI: https://doi.org/10.1049/iet-com.2019.0502

Salal, YK, Abdullaev, SM and **Kumar, M.** 2019. Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Tech*, 8(4C): 54–59.

Shrifan, NHMM, Akbar, MF and **Isa, NAM.** 2022. An adaptive outlier removal aided k-means clustering algorithm. *Journal of King Saud University-Computer and Information Sciences*, 34(8): 6365–6376. DOI: https://doi.org/10.1016/j.jksuci.2021.07.003

Subha, RP. 2019. Tree oriented association rule mining of multiple data sources. *International Journal of Enterprise Network Management*, 10(3–4): 272–279. DOI: https://doi.org/10.1504/IJENM.2019.103156

Sun, H. 2019. Study on application of data mining technology in university computer network educational administration management system. *Journal of Intelligent & Fuzzy Systems*, 37(3): 3311–3318. DOI: https://doi.org/10.3233/JIFS-179133

Thottathyl, H, Kanadam, KP and **Panchadula, RP.** 2020. Microarray breast cancer data clustering using map reduce based K-Means algorithm. *Rev. d'Intelligence Artif*, 34(6): 763–769. DOI: https://doi.org/10.18280/ria.340610

Trakunphutthirak, R and **Lee, VCS.** 2022. Application of educational data mining approach for student academic performance prediction using progressive temporal data. *Journal of Educational Computing Research*, 60(3): 742–776. DOI: https://doi.org/10.1177/07356331211048777

Wang, C and **Zheng, X.** 2020. Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. *Evolutionary Intelligence*, 13(1): 39–49. DOI: https://doi.org/10.1007/s12065-019-00234-5

Wang, L and **Soo-Jin, C.** 2021. Sustainable development of college and university education by use of data mining methods. *International Journal of Emerging Technologies in Learning* (Online), 16(5): 102. DOI: https://doi.org/10.3991/ijet.v16i05.20303

Xia, Y. 2021. Big data based research on the management system framework of ideological and political education in colleges and universities. *Journal of Intelligent & Fuzzy Systems*, 40(2): 3777–3786. DOI: https://doi.org/10.3233/JIFS-189411

Zhan, Y, Tan, KH and **Huo, B.** 2019. Bridging customer knowledge to innovative product development: a data mining approach. *International Journal of Production Research*, 57(20): 6335–6350. DOI: https://doi.org/10.1080/00207543.2019.1566662

TO CITE THIS ARTICLE:
Fang, F. 2023. A Study on the Application of Data Mining Techniques in the Management of Sustainable Education for Employment. *Data Science Journal*, 22: 23, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2023-023

Submitted: 13 February 2023
Accepted: 08 May 2023
Published: 19 July 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.