Hindawi
Security and Communication Networks
Volume 2022, Article ID 5806829, 10 pages
https://doi.org/10.1155/2022/5806829

Research Article

Data Mining Method under Model-Driven Architecture (MDA)

Jiangning Xie,[1,2] Feng Xu,[2] Zhen Li,[3] and Xueqing Li[4]

[1]*Graduate School, Shandong University, Jinan, China*
[2]*School of Management, Shandong University, Jinan, China*
[3]*Informatization Office of ShanDong University, Jinan, China*
[4]*School of Software, Shandong University, Jinan, China*

Correspondence should be addressed to Jiangning Xie; xjn@sdu.edu.cn

Received 14 January 2022; Revised 9 February 2022; Accepted 10 February 2022; Published 22 March 2022

Academic Editor: Thippa Reddy G

With the development of university information technology, how to mine and visually analyze the data of the existing separated information system will become an important research topic. The current university information system is a combination of some proprietary business systems characterized by poor data separation and storage and data analysis power. In addition, the data mining methods based on cloud computing will make customers gradually lose the ability to control the data. Because of the above problems, this paper proposes a university data mining method based on the MDA idea by constructing a data analysis and visualization framework, including multidimensional data modeling, data extraction, and data display based on visualization technology. The framework makes full use of the design idea of MDA and models multidimensional data, data extraction, and data display, respectively. The multidimensional data model module, data extraction module, and data visualization module provide efficient solutions for data analysis and visualization in universities.

1. Introduction

With the development of computer technology and network technology, various university business information management systems based on the network, including enrollment management, academic affairs management, graduation management, and financial management, have been widely used in universities. The scale and function are constantly expanding. With the continuous increase of the number of colleges and universities, information management more and more reflects its unique advantages. Many universities have gradually launched the student management information system. Functions cover student enrollment, training management, course performance management, graduation employment management, etc. Realize the standardization, informatization, and network of student management. At the same time, it has also accumulated a lot of information and data in its daily work. Also, it has the storage, backup, query, and simple statistical functions of massive data, but there are still the following problems:

First, a large amount of data is accumulated in each management stage of the system, but most data are separated into different system databases. Second, the management system realizes the management function of the data. Still, the data analysis function is relatively weak. The lack of multiangle analysis and data statistics is not enough to excavate the valuable information hidden in the massive data not to provide enough decision support for the school business managers. Third, the statistical analysis function in the management system is relatively simple, and the analysis results are primarily displayed in the form of reports and data tables, not intuitive enough. The present analysis is limited to the simple number of people, grades, courses, and so on, less to give the problems reflected by the data. It is difficult to recall the internal relationship between the data. Therefore, using statistical analysis and data mining methods to analyze the student management data from multiple angles and how to use the analysis results to provide accurate, intuitive, and good decision support for various management departments has become the focus and key of the current research.

The literatures [1, 2] proposed a data analysis method combining data mining for information management in colleges and universities, but just for a particular business problem to put forward a basic solution of ideas, lack of technical solutions for overall business data analysis. At the same time, the lack of data visual analysis function cannot provide intuitive visual analysis function. A solution that organically combines cloud computing technology and data mining is proposed in the literature [3, 4]. This solution can effectively solve the problem of the massive data and the limited computing power of the traditional data mining systems caused by the exponential growth of the data. The crow search algorithm has been successfully applied for the optimization of the data mining process, based on the characteristic of less parameter settings, easy implementation, and strong optimization capacity [5]. The performance of the system is tested through the mining and analysis of the electronic literature access log data set of college teachers and students in the library. The real-time performance and reliability of the system are verified.

For the problems of data analysis and data visualization, this paper comprehensively studies data mining [6] and data visualization technology [7, 8]. It proposes the analysis and visualization framework for college information data based on MDA [9] ideas. The framework makes full use of the design idea of MDA and models multidimensional data, data extraction, and data display, respectively. The multidimensional data model module, data extraction module, and data visualization module provide efficient solutions for data analysis and visualization in universities. Businesses and developers can quickly complete functional development for a thorough data analysis and visualization business through this framework.

2. The MDA-Based Analysis and Visualization Framework

For college students' data analysis and visualization problems, this paper designs and implements the data acquisition, analysis, and visualization framework based on MDA [10], which mainly includes three submodules of analysis data visualization modeling, data acquisition, and visualization display. The overall design structure diagram of the framework is shown in Figure 1.

In analyzing data and visualization, it is first necessary to visually model the relevant business data to build the data model. Then, data extraction and cleaning are done according to the established data model to obtain all the data sets to be displayed. Finally, the data is visually visualized and analyzed through the interactive defined display model.

2.1. Data Analysis Modeling. The data model is a typical performance of data structure. According to the needs of user-oriented, data models gradually establish different degrees of detail and refinement, which is an understanding of various degrees of abstraction in the real world. The data model in the traditional business processing system is a relational data model, which cannot effectively reflect the structure and semantic information between the data. The primary purpose of a data analysis system is to analyze operations on a particular topic, which are called facts or measures. In contrast, the various angles of the analysis are called dimensions. Therefore, a multidimensional data model is used to model the data analysis and visualization systems in this paper. The system can complete a trend analysis, a continuous time subset of data sections, and quickly create a new view representing this section.

The establishment of the multidimensional data analysis model can integrate various kinds of data details and summarize the comprehensive information to meet the needs of the decision support system. However, the establishment of the model can be effectively organized in various parts, form complete and summary data for decision analysis, and provide strong decision analysis support for the leadership decision-making layer. The logical structure design of multidimensional data model is mainly the structure design of dimension table and fact table. For the logical definition of relational patterns, the patterns should be divided according to the current implemented topics to form multiple specific dimension tables and fact tables and determine the relationship patterns of each table.

Multidimensional data models organized through the dimensional table-fact table structure form of multidimensional phenotypes can be expressed in star mode, snowflake mode, or fact constellation pattern form. Multidimensional data models often regard the data as the form of a data cube, and the data cube is defined by dimension and facts. Dimension is about the perspective or entity that an organization wants to record and collects the same class of data. This paper adopts the star mode to ensure the performance of the data query and the easy understanding of the model. Its multidimensional data model will be established as follows:

Step 1: to determine the analysis topic, assuming that the analysis topic is to analyze the teacher's performance in the past year, which is also a fact or measure in the multidimensional data model.

Step 2: to determine the analysis dimension, including scientific research dimension, teaching dimension, guidance student dimension, post-level assessment dimension, and academic reputation dimension, in which each size can be divided according to the actual situation, into smaller dimensions, to slice the data cube and other operations.

Step 3: according to the analysis of the first two steps, a multidimensional data model of the star pattern representation can be obtained.

2.2. Data Procurement. For a specific business analysis topic, after completing the creation of the multidimensional data model, the data extraction next needs to obtain the necessary data to analyze and visualize the data from different data sources. To provide clean, complete, accurate, uninformative data for the data analysis process, improve the efficiency of the analysis process, and ensure the rapid generation of the

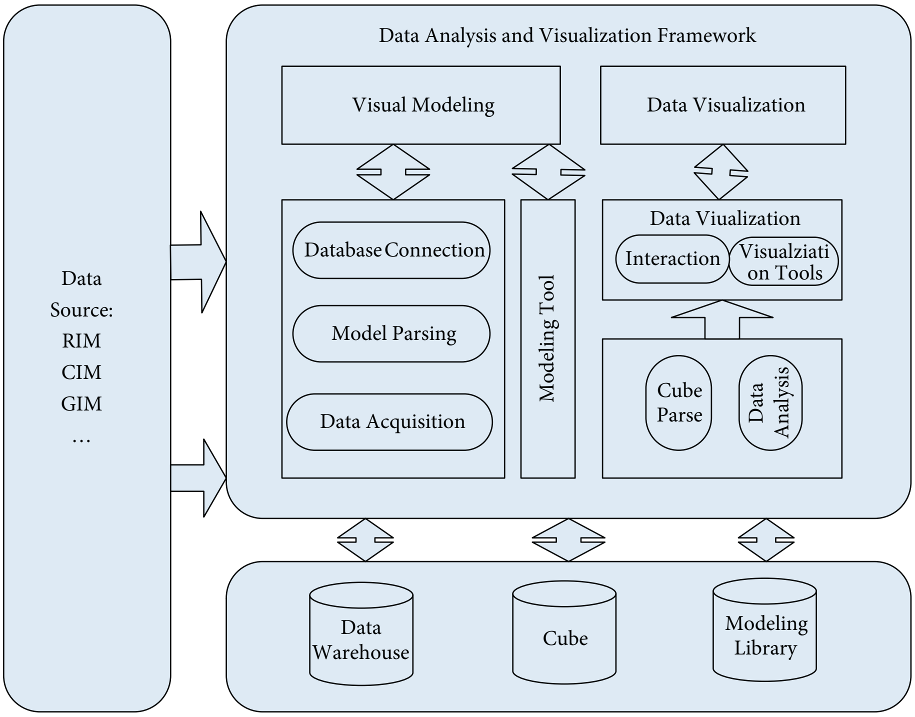

Figure 1: Data analysis and visualization framework.

presented results, this paper designs a method of data extraction, data conversion, and data loading from a data source to a cube (namely, the ETL process). The framework for the data acquisition is shown in Figure 2. In this framework, the data obtained from the source is not directly written to the cube. Instead, the data is preprocessed first, then data conversion and cleaning according to the correspondence between source and target data. Convert good data as intermediate data. The intermediate data is then stored into the cube after profound transformation.

As shown in Figure 2, the model divides the entire application system into three layers: data source layer, intermediate data layer, and multidimensional data layer. The data source layer can be divided into structured and unstructured data according to the data characteristics. This paper adopts different data acquisition methods for the above two different types of data. The paper uses traditional data extraction, data transformation, and data loading processes for structured data. This paper uses the interface-based design for unstructured data, using four typical interface methods: Web Service, intermediate library, file, TCP/UDP message transmission. Transparency to heterogeneous data acquisition is achieved by interface mode. Then, the data acquisition is completed in the transformation and loading mode of the heterogeneous data obtained in different ways.

2.3. Data Presentation. Data visualization is a theory, method, and technique for using computer graphics and image processing techniques to convert data into images or images displayed on the screen and perform interactive processing. Data visualization changes the traditional way of showing data relationships through the relationship tables, allowing people to make more intuitive and efficient observations of the relationship between data. After completing the multidimensional data modeling and data acquisition process, this paper presents the data in a visual way. The general way of data display is to convert the obtained data into a vector chart or bitmap and display the bar chart pie chart on the vector chart or bitmap. For example, SVG generates a vector map for data amount visualization. The Scalable Vector Graphics (SVG) describes a vector drawing standard for two-dimensional vector graphs in the XML language, including rectangles, circles, and polygons. It has the advantages of high graphics quality, small files, and rich performance effect, but SVG plugin must be installed at the user's browser end, which inconveniences customer browsing.

A Flex technology-based data presentation model is chosen to improve the applicability of data presentation to different architecture designs, including B/S and C/S architectures. The data display can meet the display requirements of both B/S and C/S architecture and can meet users' needs. Through the data display mode of Flex technology, the overall data display process includes data transmission, the display mode of interactive processing, and data binding. The result data is obtained from the callback mechanism of front-end Flex communication with Java. During the data conversion between Flex and java, Flex implements the data transformation and binding via a binary AMF protocol. This paper provides a visual analysis and display of data to support high-dimensional data, including pie charts, bar charts, line charts, scatter plots [11], and parallel axes [10, 12].

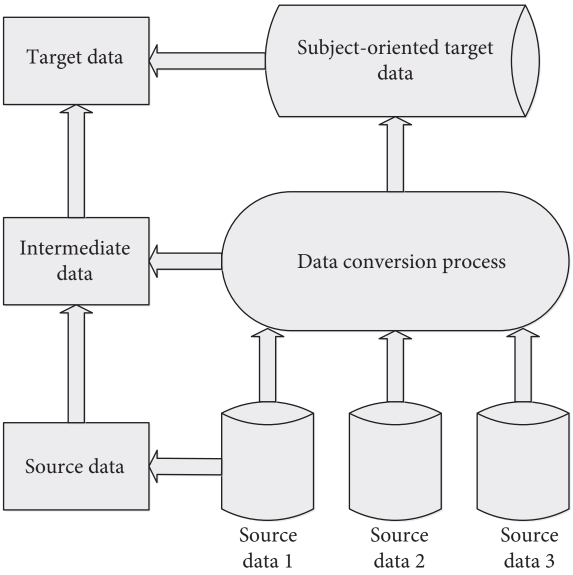

Figure 2: Data acquisition framework.

3. Example Analysis

Step 1. Determine the theme of the analysis. The theme is to analyze the school performance of the doctoral students in Shandong University to provide the school doctoral student training policy and establish an effective decision-making system. Mainly according to the daily routine of postgraduate students in school, the doctoral data is analyzed to predict the proportion of general, reasonable, and excellent academic performance. All administrative departments of the school can adopt appropriate policies to improve the research performance of doctoral students from the categories of general and good performance.

Step 2. First, obtain the behavioral data of the doctoral student performance, and then, according to the doctoral student data, determine the analysis dimension, including the scientific research dimension, the degree dimension, and the performance dimension, in which each dimension can be divided into smaller dimensions according to the actual situation. In the downward subdivided dimension, the results of various subjects to calculate the achievement dimension are used. In this example, the three main courses of doctoral students are used for analysis. The research dimensions consider published papers and influencing factors. Because of a degree dimension, it mainly evaluates the tutor evaluation results of the doctoral thesis, whether to delay defense or repeat defense.

Step 3. The model is designed from the analysis of the first two steps. In this example, the K-Means algorithm and the PCA algorithm are used to build a model framework for the cluster analysis of doctoral data. This model first clusters the doctoral student data through K-Means, after which each doctoral student corresponds to a class. Later, the doctoral student data is reduced through the PCA algorithm. The results obtained from clustering are combined with the source data and analyzed using visualization techniques.

In this example, *Python* is a programming language compatible with many platforms that support both process-oriented and object-oriented programming and include various standard libraries. The data clustering analysis is mainly used for the sklearn library and visualized to the matplotlib library.

3.1. Data Procurement. In this example, we establish excellent data standards for school data and definition of metadata, data items, data classes, and datasets. In the data standards, 15 data categories, including departmental units, student management, teaching management, staff management, and scientific research management, are established, and each data class contains its respective subclass. The data interaction is related to huge amount of data of undergraduate schools, graduate schools, and colleges through the data standards. We extracted the daily behavior performance data with doctoral students.

Through the school system, we collect data on doctoral students' daily behavior and scientific research performance and use the traditional data extraction, data transformation, and data loading process for processing. Scientific research information, degree information, and results are all structured data.

The collected data are 3579 pieces of doctoral data. The doctoral degree types include general master and postgraduate degrees, general professional doctors, general doctors, and general direct degrees. The disciplines come from medicine, science, law, engineering, and other fields. In the college distribution, these doctors come from many colleges, including the School of Pharmacy, the Law School, the Business School, and the Institute of Economics. This example mainly considers information from 11 dimensions to analyze it.

The dimensions that assess doctoral performance are 11. These mainly include the following:

(1) Degree information is divided into MidCheck, answer, comment1, comment2, comment3, commentNum. It represents the degree information in the doctoral thesis review, and the larger the number of comments made by the reviewing supervisor, the worse the doctoral student's performance.

(2) Achievement dimension information is divided into Course1, course2, course3. They represent the results of three doctoral subjects. The higher the performance, the better the doctoral student performs.

(3) The scientific research dimension information is as follows: DisserNum, impactFactor. It represents the sum of the number of published papers and the influence factors of the papers, respectively. The larger the number, the better the doctoral student's performance.

3.2. K-Means Model. We send the extracted doctoral data into the cluster model for clustering to obtain the label values for their categories when analyzing the doctoral data. We use the K-Means algorithm [13] for clustering statistics. The K-Means showed fast convergence, excellent clustering effect, and strong interpretable algorithm, so we used K-Means as a model for clustering. In the model, we randomly divided the data into K groups and selected K objects as the initial clustering center. The distance between each data point and each cluster center is calculated after being assigned to the nearest cluster center. When each data point is assigned to the corresponding data center, the clustered cluster center is recomputed based on the existing cluster data point distribution. This step is repeated continuously until all points are clustered. The clustering result has the smallest sum of error.

In the K-Means, the distance measure used is the square of the Euclidean distance:

$$\begin{aligned} d(x, y)^2 &= \sum_{i=1}^{n} (x_i - y_i)^2 \\ &= \|x - y\|_2^2, \end{aligned} \tag{1}$$

where x, y represent two different samples, and n represents the dimension of the sample. The problem with the Euclidean algorithm is that the sum of squares of error (SSE) in the cluster is minimized with the following formula:

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{m} w^{(i,j)} = \left\| x^{(i)} - \mu^{(j)} \right\|_2^2, \tag{2}$$

where $\mu^{(j)}$ represents the central point of the cluster j.

We determine the K cluster using the elbow method. As the number of clusters, K, increases, the sample division will be gradually detailed, the aggregation degree of each class of clusters will gradually increase, and the resulting sum of errors (SSE) will gradually decrease. When K is smaller than the real number of clusters, the SSE will drop greatly because the K greatly increases the aggregation of each cluster. When the K reaches the real cluster number, the aggregation obtained by the K will decrease rapidly, so the decline of SSE decreases sharply. Then, the curve will gradually flatten as the K value increases. The curve is similar to the elbow shape, with the corresponding K value as the true cluster number of the data.

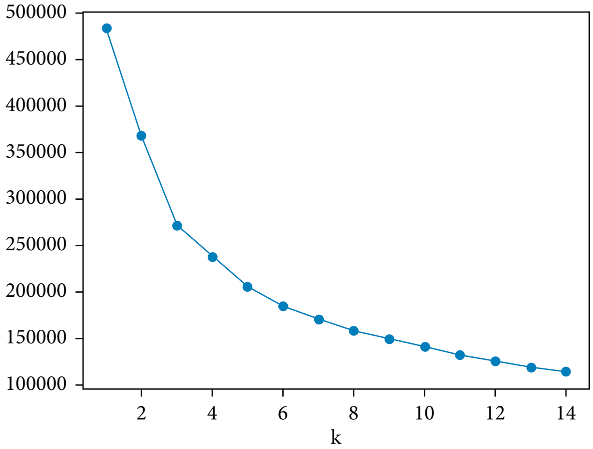

Figure 3: Elbow method image of doctoral student data.

Classification images of doctoral students are drawn by elbow method, as shown in Figure 3.

The horizontal axis of the image is K, and the vertical axis is SSE. The image shows that the inflection point is $k = 3$ when the clustering class is 3.

After determining the K value, we set the K to 3 and send it to the K-Means model for clustering operations, in order to obtain the best clustering data.

The algorithm steps for K-Means are in Algorithms 1 and 2.

α_j First, the initialized k samples are selected as the initial cluster center $\alpha = \alpha_1, \alpha_2, \ldots \alpha_k$.

(1) For each sample x_i in the dataset, its distance to the k cluster center is calculated and divided into the class corresponding to the cluster center with the smallest distance.

(2) For each class α_j, recalculate its cluster center $\alpha_j = 1/|c_i|\sum_{x\epsilon c_i} x$ (the center of mass of all samples belonging to the class) for each class;

(3) Repeat the above 1 and 2 steps until some abort condition (number of iterations, minimum error change, etc.) is reached.

After completing the multidimensional data modeling and data acquisition process, we obtained 3579 doctoral data sample labels and clustering results of the data, followed by dimensionality reduction and presentation of the data in a visual manner.

3.3. PCA Dimension Reduction. The main idea of PCA [14] is to map n-dimensional features to k-dimensions. This k-dimensional feature is an entirely new orthogonal feature, called the principal component, and is a k-dimensional

```
$n \times m\mathrm{Arr}_{n\times m}$ Input: dimension array, iteration t
Output: The label value $\mathrm{Label}_n$
Begin
    Automatic_Random_Generate() $\mathrm{point}_K$
    while(t)
      for(int $i = 0$; $i < n$; $i$++)
        for(int $j = 0$; $j < k$; $j$++)
          Calculate_Distance() $\mathrm{Arr}_i\mathrm{point}_j$
      for(int $i = 0$; $i < k$; $i$++)
        Find_All_Data_Points_belong_Cluster() $\mathrm{Arr}_{n\times m}\mathrm{point}_K$
        Modify_Coordinate_to_Center_Coo_Points() $\mathrm{Arr}_{n\times m}\mathrm{point}_K$
End
```

Algorithm 1: *K*-Means clustering.

feature reconstructed based on the original n-dimensional feature. PCA's work is to find a set of mutually orthogonal axes from the original space sequentially, and the selection of the new axes is closely related to the data itself. The first new axis selection is the direction of the largest variance in the original data. The second new axis selection is the plane orthogonal to the first axis and the largest in the plane orthogonal to the 1 and 2 axes. By this, in turn, n such axes can be obtained. With the new axes obtained this way, we find that most of the variances are contained in the preceding k axes, and the latter axis contains a variance of almost 0. Therefore, we can ignore the remaining axes and retain only the first k axes containing most of the variance. We retain only the dimensionality features containing most of the variance in the doctoral data, while ignoring the feature dimension containing the variance of almost 0, achieving the dimensionality reduction of the data features.

PCA algorithm based on eigenvalue decomposition covariance matrix is in Algorithm 2.

We transform the doctoral data into m n-dimensional vectors $X_{m\times n}$ for and performed zero-mean value (the average of this column) for it. The covariance matrix C is found, which further obtains its eigenvalue and eigenvector A. The eigenvectors λ form the matrix with the corresponding eigenvalues from large to small and then take the first k to form the matrix P. Y=PM is a matrix of a new k-dimensional size.

By calculating the covariance matrix of the data matrix, the eigenvector of the eigenvalues of the covariance matrix is obtained, and the matrix composed of the eigenvectors corresponding to the k features with the largest eigenvalue is selected. Transfer data matrix into a new space to achieve dimensionality reduction of data features. The covariance is

$$\begin{aligned}\mathrm{COV}(X,Y) &= E[(X-E(Y))(Y-E(Y))]\\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}).\end{aligned} \tag{3}$$

For the n-dimensional matrix, the covariance matrix is

$$C = \left(c_{ij}\right)_{n\times m} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}, \tag{4}$$

$c_{ij} = \mathrm{Cov}(X_i, X_j)$, where $i, j = 1,2,3$, ..., n.

The divergence matrix is defined as

$$S = \sum_{k=1}^{n}(x_k - m)(x_k - m)^T, \tag{5}$$

where m is the average vector: $m = 1/n\sum_{k=1}^{n} x_k$.

Vector v is the eigenvector of matrix A and can be expressed in the following form:

$$Av = \lambda v. \tag{6}$$

Among them, λ is the eigenvalue corresponding to the eigenvector v, and a set of eigenvectors of the matrix is a set of orthogonal vectors.

For matrix A, there is a set of eigenvectors v, which are orthogonalized to obtain a set of orthogonal unit vectors. Eigenvalue decomposition decomposes the matrix A into the following formula:

$$A = Q\sum Q^{-1}. \tag{7}$$

Among them, Q is a matrix composed of the eigenvectors of the matrix A, and $\sum$ is a diagonal array, and the elements on the diagonal are the eigenvalues.

In this example, we use the PCA algorithm and divide it into two dimension reduction methods:

(1) We reduce the degree information midCheck, answer, comment1, comment2, comment3, commentNum and research dimension disserNum, impactFactor to 1 dimension information, representing the degree level and research level of doctoral students. The achievement dimension information course1, course2, and course3 are reduced to 1-dimension information, which indicates the performance level. After the dimensionality reduction

```
Input: $m$ n-dimensional vectors $X_{m\times n} = \{x_1, x_2, x_3, \ldots, x_m\}$
Output: $k$ n-dimensional vectors $Y_{k\times n} = \{y_1, y_2, y_3, \ldots, y_k\}$
begin
      for($i = 0$; $i < n$; $i$++)
         $\mathrm{M}_i$ = Minus_Average $x_i$
   $C = 1/mMM^T$
   $\lambda, A$ = Solve_covariance(C)
   $P$ = Array_Dwindle(A[1:k])
   $Y$ = PM
end
```

Algorithm 2: PCA algorithm.

through the PCA algorithm model, we reduce the 11-dimensional intake to 2.

(2) We reduce the degree information midCheck, answer, comment1, comment2, comment3, and commentNum to 1-dimension information representing the degree level of a doctoral student. The achievement dimension information course1, course2, and course3 are reduced to 1-dimension information, which indicates the performance level. Reduce the scientific research dimension disserNum and impactFactor to 1-dimensional details, representing the scientific research level. After the dimensionality reduction through the PCA algorithm model, we reduce the 11-dimensional intake to 3.

After the doctoral data underwent PCA dimension reduction, we combined the source data for analysis and presented them with visualization techniques.

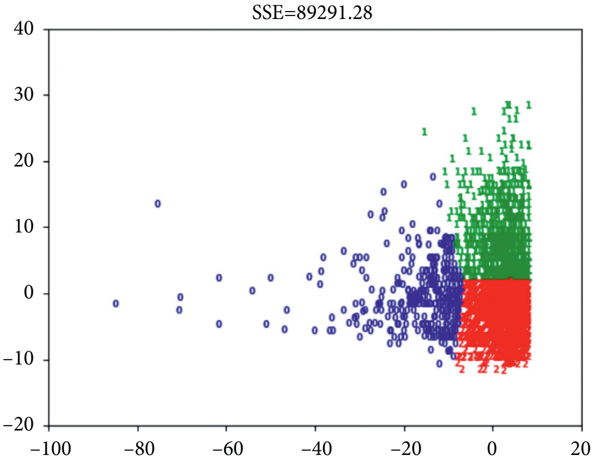

Figure 4: Doctoral cluster image of doctoral students, horizontal axis represents research degree information and vertical axis represents achievement information.

3.4. Data Display. After the model operation, we can get the picture effect of the doctoral student data clustering. The 2 *D* clustering effect is shown below, and the total number of doctoral students is 3579, including 2052 in class 0 (blue), 1179 in class 1 (green), and 348 in class 2 (red). The drawing invokes the matplotlib graphics library for the python programming language. matplotlib is a library dedicated to developing 2D charts and 3D charts, realizing data visualization gradually and interactively, strong control over image elements, and output multiple formats including PNG, PDF, SVG, and EPS, as shown in Figure 4.

From the results of data clustering image classification, combined with the analysis of Tables 1 and 2, compared with the mean value, class 0 (blue) doctora *l* students perform the best, and their scientific research, degree, and performance are relatively excellent. The number of papers (disserNum) and paper impact factors (impactFactor) exceeded the other two doctoral students. The impact factors are nearly five times more than the two different categories. Among the achievement items, class 0 students had the highest grades in course 1 and course 3, with course 2 at the middle level.

There are some problems in class 2 (red) doctoral students—their scientific research ability performance in general. The number of papers and paper influence factors is relatively small, but their performance is good, the best in course2 and course3. Their degree performance is in the middle level. Class 1 (green) doctoral students have the least number of scientific research papers. Still, the impact of the papers is relatively high, with the worst performance, and their degree performance is in the middle level.

Based on the above data, we continue to explore the classification results of $k = 4$ with $k = 5$. After data analysis, we found that the subdivided types of doctoral students are more detailed in terms of scientific research, degree, and performance.

In the $k = 4$ classification, 9 doctoral students with extra categories of 2 (red part) in Figure 5(a) have better degree information, good grades, the best number of published papers, and the highest impact factors, because fused Ph.D. data from class 0 in $k = 3$. In the classification of $k = 5$, according to the source data analysis, the influence factors of the doctoral data with class 4 in Figure 5(b) papers are significant, but their academic performance is low.

After collecting the above data analysis results, we further explored that we standardize the doctoral data and processed the standardized data through K-Means and PCA. Then, two-dimensional images are then obtained, as shown in Figure 6. From the data distribution perspective, the data distribution after the standardization is somewhat more uniform.

Table 1: PhD student data $K = 3$ cluster research and achievement mean.

Classify	Disser Num	Impact factor	Course 1	Course 2	Course 3
0 class	4.6041	24.2042	89.7470	88.5654	87.2619
1 class	2.0813	5.4582	88.1742	80.8182	86.1568
2 class	2.2335	5.0030	88.3930	91.9798	88.1305

Table 2: PhD student data $K = 3$ cluster degree information mean.

Classify	MidCheck	Answer	Comment 1	Comment 2	Comment 3	Comment Num
0 class	1.2321	2.7351	1.9375	1.7202	1.3184	7.9315
1 class	1.0871	2.4655	1.9145	1.6506	1.2663	7.7668
2 class	1.1133	2.4170	1.9156	1.6496	1.2654	7.7894

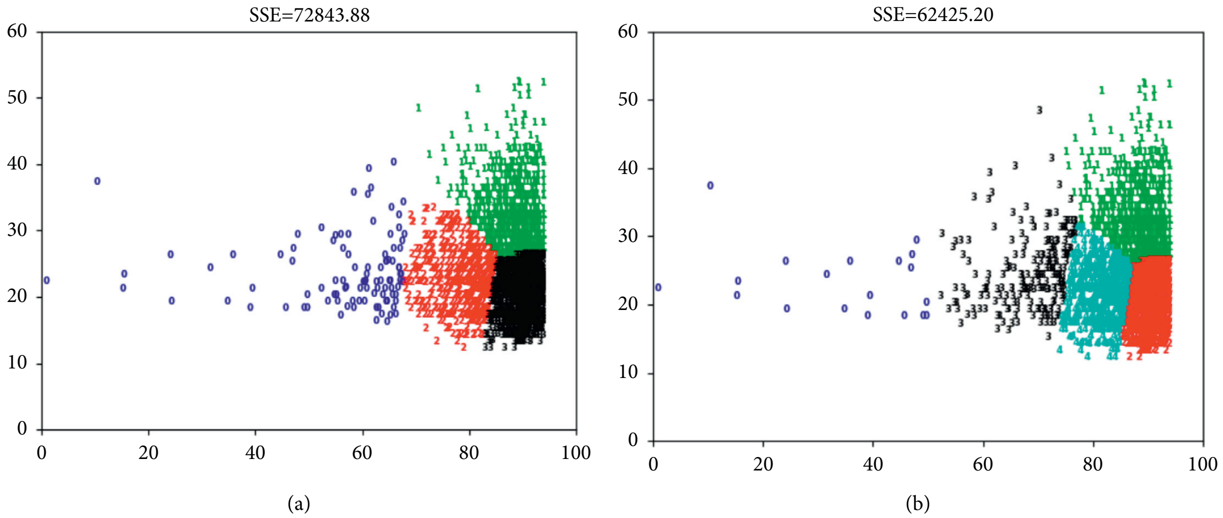

Figure 5: (a) PhD student cluster images, $k = 4$. The horizontal axis represents scientific research degree information, and the vertical axis represents achievement information. (b) PhD student cluster images, $k = 5$. The horizontal axis represents the scientific research degree information, and the vertical axis represents the achievement information.

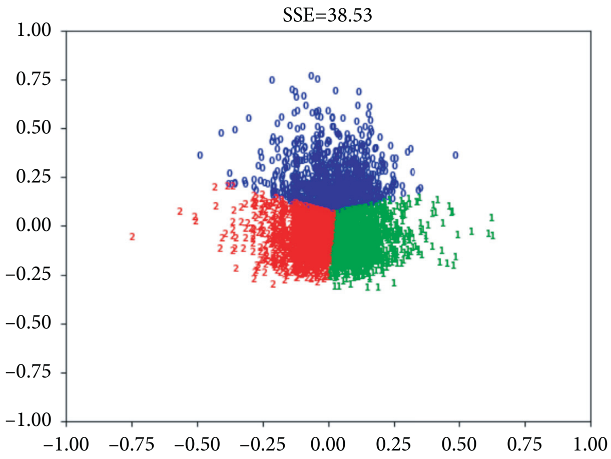

Figure 6: Cluster images of doctoral students after standardization: horizontal axis represents scientific research degree information, and vertical axis represents achievement information.

We analyze the data and averaged the clustering results according to the data clustering results. It can be found from Table 3 that class 0 (blue) doctoral students are at the middle level. Their grades are at the downstream level. Class 1 (green) doctoral students have the best scientific research, degree, and performance. These kinds of students are

Table 3: Data mean values after normalization.

Average value	Scientific research degree	Mark
0 class	0.350780663	0.654025096
1 class	0.407664049	0.784348448
2 class	0.302289483	0.773064877

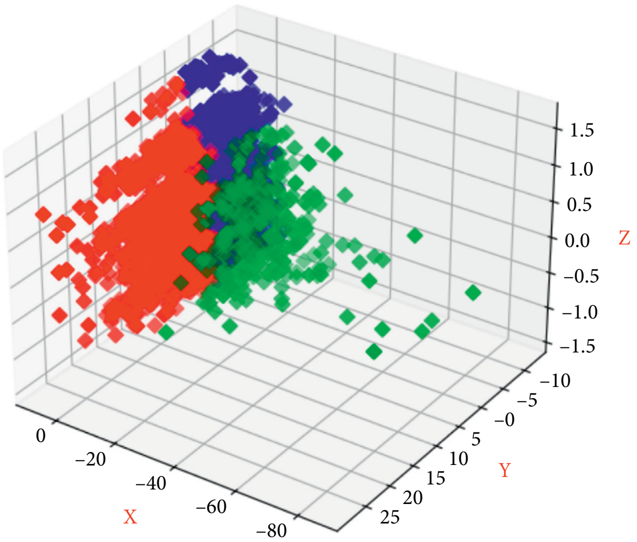

Figure 7: 3 D doctoral student cluster image: x-axis represents scientific research information, y-axis represents achievement information, and z-axis represents degree information.

relatively excellent. Class 2 (red) has a low scientific research and degree life, but their grades are in the middle level.

We further expand based on the analysis of the above 2 D data. When we use PCA to reduce 11 D doctoral information to 3 D, we can build 3 D graphics, as shown in Figure 7.

From the analysis of data clustering results combined with doctoral performance data, we see that green doctoral students have the best performance. Their scientific research, degree, and performance are relatively excellent. Red doctoral students have general scientific research ability. Their performance is the worst, and their degree is middle. Blue doctoral students have general scientific research ability, good performance, and degree performance.

From the above analysis, it is learned that appropriate data analysis applied to doctoral student performance can effectively extract effective information from a large amount of data. It can be used for the school management decision-making process. Statistical analysis of doctoral data helps provide adaptive learning guidance. Preserving the students' performance and behavior in advance will help the relevant departments of the school to take appropriate adjustment measures to cultivate talents better and build a talent system.

4. Conclusion

Data analysis and visualization technology is an emerging and promising research field in university information management. With the development of university information technology, how to mine and visually analyze the data of the existing separated information system will become an important research topic.

This paper presents a complete set of solutions based on MDA ideas for analyzing and visualizing information data in universities. The proposed framework includes multidimensional data modeling and analysis modules, data extraction and cleaning modules, and data display modules based on data visualization techniques.

Through this framework, the business analysis and developers can quickly conduct the data visualization business's modeling analysis and programming implementation. In the existing framework, the interface is provided for data mining analysis. Moreover, the data mining algorithms and multidimensional data visualization techniques will be deeply studied and applied to the existing data analysis and visualization frameworks in future work.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The authors received no financial support for the research, authorship, and/or publication of this article.

References

[1] Y. Xu and D. Xiong, "Research on the information construction of college student management in the era of big data," *Journal of Lanzhou Education College*, vol. 34, no. 1, p. 3, 2018.

[2] Q. Zhang and F. Rui, "Application of data mining in university information management," *Data Mining*, vol. 9, no. 1, p. 7, 2019.

[3] R. Zhong and H. Wang, "Specific data query technology in the university cloud computing management system based on data mining," *Modern Electronic technology*, vol. 41, no. 2, p. 3, 2018.

[4] G. Shen, "Application of big data analysis in smart education in universities," *Modern Electronics Technology*, vol. 42, no. 4, pp. 105–108, 2019.

[5] C. Qu and Y. Fu, "Crow search algorithm based on neighborhood search of non-inferior solution set," *IEEE Access*, vol. 7, 2019.

[6] A. R. Raut and S. P. Khandait, "Review on data mining techniques in wireless sensor networks," in *Proceedings of the IEEE Sponsored 2nd International Conference on Electronics and Communication System(ICECS 2015)*, February 2015.

[7] D. M. Schug, P. H. Taylor, S. Iudicello, and J. H. Swasey, "Using online data visualization and analysis to facilitate public involvement in management of catch share programs," *Marine Policy*, vol. 122, 2020.

[8] G. Ralitza, B. Eugenia, S. Patricia et al., “Data visualization tools of tobacco product use patterns, transitions and sex differences in the PATH youth data,” *Nicotine & Tobacco Research*, vol. 22, no. 10, pp. 1901–1908, 2020.

[9] B.-K Park and W.-S. Jang, “MDA(Model driven architecture),” *Journal of Platform Technology*, vol. 7, 2019.

[10] L. Lu, W. Wang, and Z. Tan, “Double-arc parallel coordinates and its axes re-ordering methods,” *Mobile Networks and Applications*, vol. 25, no. 4, pp. 1376–1391, 2020.

[11] S. Cao, Y. Zeng, S. Yang, and S. Cao, “Research on Python data visualization technology,” *Journal of Physics: Conference Series*, vol. 1757, no. 1, pp. 012122–012128, 2021.

[12] G. Richer, J. Sansen, F. Lalanne, D. Auber, and R. Bourqui, “HiePaCo: scalable hierarchical exploration in abstract parallel coordinates under budget constraints,” *Big Data Research*, vol. 17, no. 8, 2019.

[13] J. Li, S. Xu, Wan Can, Y. Lu, and S. Wang, “Analysis of power load characteristics based on the adaptive k-means + algorithm,” *China Southern Power Grid Technology*, vol. 13, no. 2, p. 7, 2019.

[14] J. Xie, X. Li, L. Wang, and Y. Niu, “A MDA-based campus data analysis and visualization framework,” in *Proceedings of the ETCS ’11: Proceedings of the 2011 Third International Workshop on Education Technology and Computer Science - Volume International Journal of Education and Management Engineering*, vol. 2, no. 10, Washingdon, D C USA, March 2011.