Hindawi
Mobile Information Systems
Volume 2023, Article ID 9871487, 1 page
https://doi.org/10.1155/2023/9871487

Retraction

Retracted: Research on the Application of Data Mining Technology in the Analysis of College Students' Sports Psychology

Mobile Information Systems

Received 31 October 2023; Accepted 31 October 2023; Published 1 November 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] S. Hou, "Research on the Application of Data Mining Technology in the Analysis of College Students' Sports Psychology," *Mobile Information Systems*, vol. 2021, Article ID 6529174, 7 pages, 2021.

Hindawi
Mobile Information Systems
Volume 2021, Article ID 6529174, 7 pages
https://doi.org/10.1155/2021/6529174

Research Article

Research on the Application of Data Mining Technology in the Analysis of College Students' Sports Psychology

Shujun Hou

Sangmyung University, Seoul, Republic of Korea

Correspondence should be addressed to Shujun Hou; qwerty21232021@163.com

Received 20 August 2021; Revised 22 September 2021; Accepted 15 October 2021; Published 23 November 2021

Academic Editor: Omar Cheikhrouhou

The advent of the information age has changed every existing career and revolutionized most if not all fields, notwithstanding many benefits that came along with it. There has been an exponential rise in information and, alongside it, an increase in data. Data centers have erupted with details as the number of rows in databases grows by the day. The use of technology has nevertheless become essential in many company models and organizations, warranting its usage in virtually every channel. College physical education and sports are not an exception as students studying such subjects are skyrocketing. As the information is getting more complex, improved methods are needed to research and analyze data. Fortunately, data mining has come to the rescue. Data mining is a collection of analytical methods and procedures used exclusively for the sake of data extraction. It may be used to analyze features and trends from vast quantities of data. The objective of this study is to explore the use of data mining technologies in the analysis of college students' sports psychology. This study uses clustering methods for the examination of sports psychology. We utilize three clustering methods for this aim: expectation-maximization (EM) algorithm, k-means, COBWEB, density-based clustering of applications with noise (DBSCAN), and agglomerative hierarchal clustering algorithms. We perform our forecasts based on various metrics combined with the past outcomes of college sports using these methods. In contrast to conventional data research and analysis techniques, our approaches have relatively high prediction accuracy as far as college athletics is concerned.

1. Introduction

Data mining is the "extraction" from an extensive dataset containing confidential information. Datasets have hidden data related to the features and trends in various datasets that may be "mined" by data mining methods. In data, there is essential information concealed. This knowledge is frequently buried and unused since the underlying data are produced more quickly than they can be analyzed and made sense of. Individuals or organizations with limited resources–particularly technical—will find it almost difficult to locate and get any insight from the data.

The term "data mining" refers to a set of tools and methods for "extracting" or "mining" information from vast quantities of data [1]. It is all about identifying patterns and connections in data that may lead to new understanding. Furthermore, these connections may serve as forecasters of future events.

Data mining's significance has been demonstrated for commercial applications, criminal investigations, biomedicine [2], and more recently, counter-terrorism [3–5]. For example, most merchants utilize data mining techniques to discover consumer purchasing trends. https://www.Amazon.com analyzes purchase history to offer product suggestions to customers. Data mining may be used wherever there is a large amount of data that needs to be explored.

Machine learning (ML), mathematical algorithms, and statistical models are a part of data mining technologies. In many areas, such as companies, financial institutions, and governmental organizations, ML tools were extensively utilized. To name just a few, RSA for secure information transfer, mathematical tools, and methods from such fields as a set, graph, and number theory have been used for public-key cryptosystems. Some of the techniques that are extremely important in data mining are the patterns of the

tracking of aberrations in datasets, which are used at regular intervals; classification of data attributes to indistinguishable categories to derive or use further information for the same purpose, clustering the grouping of concerning data chunks. Figure 1 is a representation of the classification versus clustering presentation.

Figure 1 indicates the differences between the clustering and classification techniques of data mining. Classification techniques have predefined outputs, whereas clustering creates subgroups based on the characteristics of the datasets.

This research mainly deals with the clustering technique and its algorithms to investigate the application of the data mining technology in sports in higher institutions of learning in China. Clustering is slightly similar to classification as they are both employed in pattern recognition in ML. However, classification uses predefined classes, whereas clustering identifies similarities between objects and "clusters" based on those similarities. Table 1 indicates the main differences between classification and clustering algorithms.

The single-phase and low complexities make clustering easier to employ in studying college data mining applications, which is why we selected it for this research.

1.1. k-Means Algorithm. k-means is an iterative algorithm that partitions the dataset into distinct predefined and nonoverlapping clusters. In this algorithm, each data point can only belong to a single group. It tries to keep intracluster data points as close as possible while maintaining the clusters as distinct (far) as possible. By assigning data points to a cluster, it can maintain the sum of the cluster's centroid (i.e., the arithmetic mean of all data points belonging to that cluster) and squared distance between the data points.. The lower the variance among groups, the more homogenized (alike) the data points within that cluster are. The algorithm first works by the number of clusters K; centroids are then initialized instead by shuffling the datasets, after which a random selection of K data points is conducted without replacement. Finally, the processes are iterated until there is no further change in the centroids. Computations are performed for the squared distance between the centroids and data points and assigned to the closest centroid. The cluster's centroids are computed by averaging all the data points belonging to each group. It employs the EM methodology where the E-step assigns the data points to the closest clusters, while the M-step is the computation process of the centroid for each set. Using the k-means cluster analysis aims to organize psychologically relevant parameters of expectations based on individual elements into statistically homogenous cluster groupings [2].

Figure 2 shows datasets before and after the k-means clustering algorithm application. Datasets with similar characteristics are "clustered"/grouped in the same cluster.

A significant property of the k-means cluster analysis is that discrepancy between the clusters is reduced due to optimization. Objects in the same group become insignificant, but differences across clusters are noteworthy. Identifying the first cluster centers is necessary to discover minor discrepancies between the indicators under consideration in one group [2, 3]. A study by [6–12] indicated a k-means clustering algorithm's merits over density-based clustering. It is because the density-based clustering does not consider all the data points informing the clusters. The study was conducted using R studio and R programming language to investigate athletes during training sessions. The research results are validated using the k-means algorithm in American football sports over the other traditional approaches [13]. Figure 2 shows the data before and after the application of k-means clustering.

1.2. Expectation-Maximization Algorithm. It is an optimization procedure like the gradient-descent algorithm with the advantage that updates can be computed analytically in many circumstances. Its flexibility also places it at a vantage position as compared to the other optimization techniques. The EM algorithm computes a close approximation of the optimal parameters effectively. After the process, a data curve is assigned to the cluster from which it is most likely to originate [14]. It is mainly used on incomplete data or one with missing data points or latent variables. Figure 3 indicates the EM algorithm on a Gaussian curve.

Figure 3 shows a Gaussian curve of density against value for the expectation-maximization algorithm.

1.3. DBS. It is a density-based clustering algorithm and slightly similar to the mean-shift algorithm. Its mechanism is to locate high-density regions separated from one another by low-density regions. It checks the epsilon ε (local expanding cluster radius) value and groups the closely fitted data points together. It is not a prerequisite to defining the number of clusters; however, the algorithm can identify outliers as noise. Notwithstanding the merits of DBSCAN, it works poorly with groups of varying densities [15]. It might also not work well with high-dimensional data due to the limitations of calculating the epsilon value. Figure 4 is a representation of the DBSCAN algorithm.

Figure 4 shows the core points, noise points, and border points of the DBSCAN clustering technique. The epsilon value is also indicated as the cluster radius.

1.4. COBWEB. It generates the hierarchical clustering. The clusters are probabilistically described. A classification tree's nodes represent classes (concepts) and are labeled with a probabilistic idea that aggregates the attribute-value distributions of objects categorized under the node [16].

Figure 5 shows tree nodes with their respective attribute values.

This classification tree may be used to forecast the new object's missing properties or class. Figure 5 is the representation of the COBWEB algorithm.

1.5. Hierarchical Clustering. It consists of two major groups, the bottom-up and the top-down (divisive). Each cluster in its cluster starts from the top-down. As it falls, the cluster pairs are combined (agglomerated) consecutively [17]. By

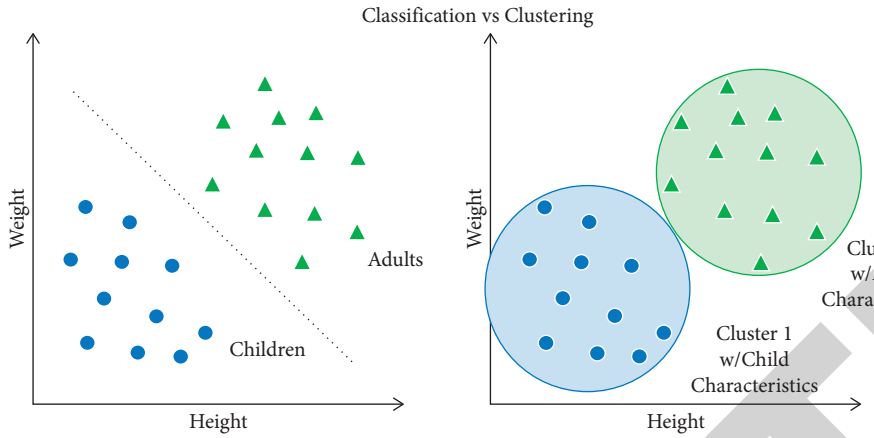

Figure 1: Classification vs. clustering.

Table 1: Differences between classification and clustering data mining techniques.

No.	Classification	Clustering
1	Labeled input data	Unlabeled input data
2	Predetermined output	Unknown output
3	Employs supervised learning	Uses unsupervised learning
4	Trained datasets are used to produce the different classifications	Prepared datasets not used to create the clusters
5	More complex	Less complex
6	Two-phase	Single-phase

contrast, the bottom-up begins as a single cluster divided recursively by moving the hierarchy down. A tree is used to illustrate the order of the clusters, and the distance measure was used to evaluate the data points. The clusters may be merged or divided based on the distance between the data points. In contrast to DBSCAN, it is not necessary to define the number of clusters.

2. Method

2.1. Statement of the Problem. Sports psychology is a popular topic in college sports and social studies classes. In most instances, descriptive research will be performed with sports clubs differentiated based on structural factors such as size, age, and the number of sports items provided. However, since sports psychology is highly varied and includes many other features, these factors are not always the most apparent. As a result, data mining is being utilized to perform more exploratory studies on the four major problems faced by sports teams. Retaining and recruiting volunteers and attracting young athletes, coaches, and members are among the subjects covered. Thematic analysis is based on the 2007/2008 Sports Development Report ($n = 13{,}068$). For each of the four questions, the decision tree is estimated. The findings indicate that, in addition to the club's size, the percentage of members who participate in social activities and the types of sports offered are significant determinants of the severity of these issues. In football, tennis, shooting, and equestrian teams, specific problems are more prevalent. Future research may be linked to comparing data mining findings in the study of Korean University sports groups.

2.2. Sampling and Study Design. College sports performance data can originate from a variety of sources. In-house statistics are the most commonly used approach. During the test, the sports' participants pedaled by increasing their power based on incremental load steps. The trainer determines the load. A spirometer and a collection of sensors gather physiological information and study their body's response to the increased effort. Each test delivers two types of information: factual data and dynamic data. Accurate information describes specific details on the user (e.g., sex, age, and weight) and information about the athletic activity, such as the test length. They give the context required to describe athletic performance and compare athletes adequately. In reality, age, gender, and weight variations between two athletes impact how their performances are evaluated and compared. The tested athletes' age, BMI, and BSA are the accurate data in this research. The body mass index (BMI) is a metric that compares a person's weight and height. Table 2 indicates some of the metrics used in the experiment.

The different heterogeneous physiological signals acquired using sensors, such as heart rate, ventilation, oxygen consumption, carbon dioxide, and oxygen concentration, provide dynamic data. Physiological signals are captured during the test, and each sample provides a snapshot of the observed athlete's state. A descriptive study was conducted in association with the NCAA basketball program from China. The study was conducted in China college institutions, where a random sampling process was completed. To track the performance of the college basketball team, Catapult Sports OPTIMEYE S5 was worn by all the participants

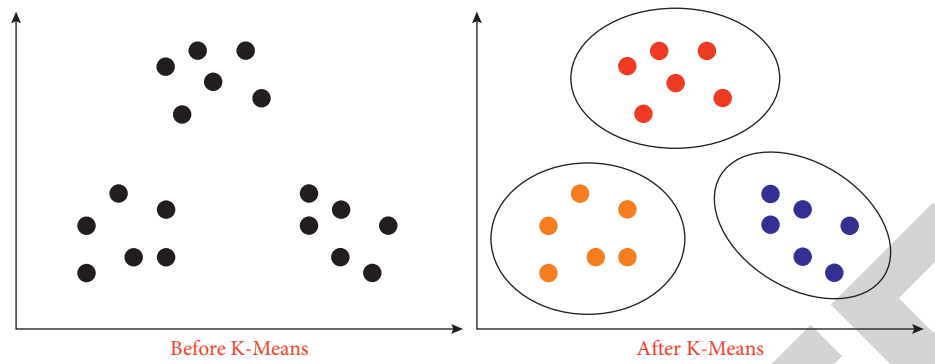

Figure 2: k-means clustering process.

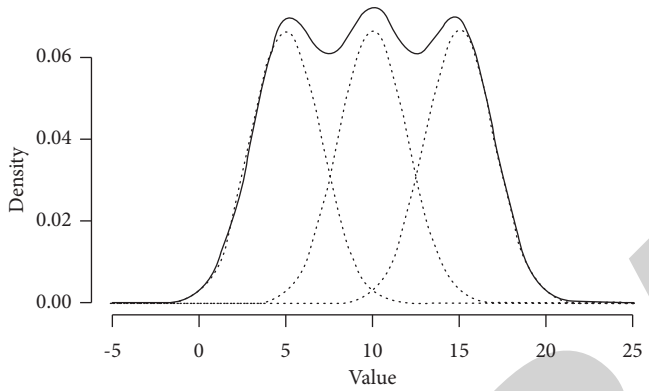

Figure 3: Representation of the expectation-maximization algorithm on a Gaussian curve.

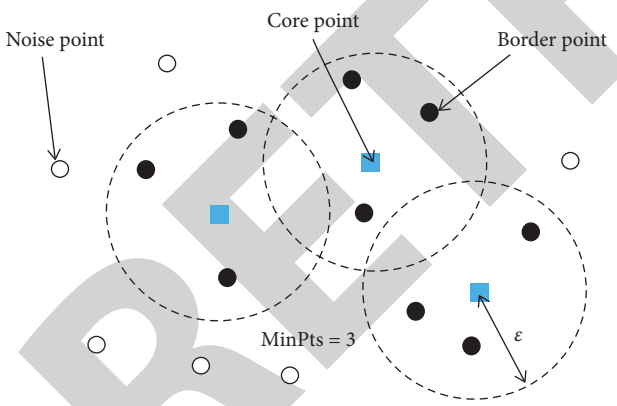

Figure 4: DBSCAN representation.

[1]. This device is designed to be worn on the upper-back side of the player's shoulder pads and to communicate wirelessly via GPS. The player's movements are captured via the GPS with an accuracy of 3 meters. The participants were of the male gender for consistency of the results. These participants were of an average height of 1.80 meters and had an average weight of 80.29 kg.

2.3. Data Preprocessing. Data cleaning and transformation were conducted on the collected data. Data cleaning was performed to remove missing and irrelevant data points. For missing data, the "ignore tuple" technique was employed. The method was selected due to the large dataset used in the research. A combination of binning and regression methods was used. For the binning process, the data were first sorted and then divided into portions of equal size. Each segment was handled separately afterward. Some data segments were replaced with their mean or boundary values. For the regression method, a regression function was used to perform the fitting process. Data transformation which converts data into a suitable form for data mining was conducted [18]. Dimensionality reduction was shown on the data to reduce the data size by encoding mechanisms. Both principal component analysis (PCA) and wavelet transforms were conducted.

2.4. Data Analysis

2.4.1. k-Means Clustering. Unlike DBSCAN and hierarchical clustering techniques, the k-means algorithm has a prerequisite to determine the number of clusters. For this process, the elbow method was performed where we parsed multiple values for k and calculated the sum of the inside of squares (WSS). We used Pandas, Numpy, Sklearn, Seaborn, and Matplotlib in k-means clustering. Using the blob data structure, we determined the shape of the data. Figure 6 is the python code used for this process.

The elbow process was able to find the optimum number of clusters using the code in Figure 7.

Figure 8 is a graphical representation of the elbow curve generated from the process.

Figure 8 indicates a graph generated after elbow process application.

We then utilize within-sum-of-squares to determine the optimal number of clusters generated for a particular dataset. The total of squared distances between each member of the cluster and its centroid is defined as the inside the sum of squares (WSS) as per the following equation [19]:

$$\mathrm{WSS} = \sum_{j=1}^{m} \left(x_j - c_j\right)^2, \tag{1}$$

where x_j is the datapoint and c_j is the point closest to the centroid.

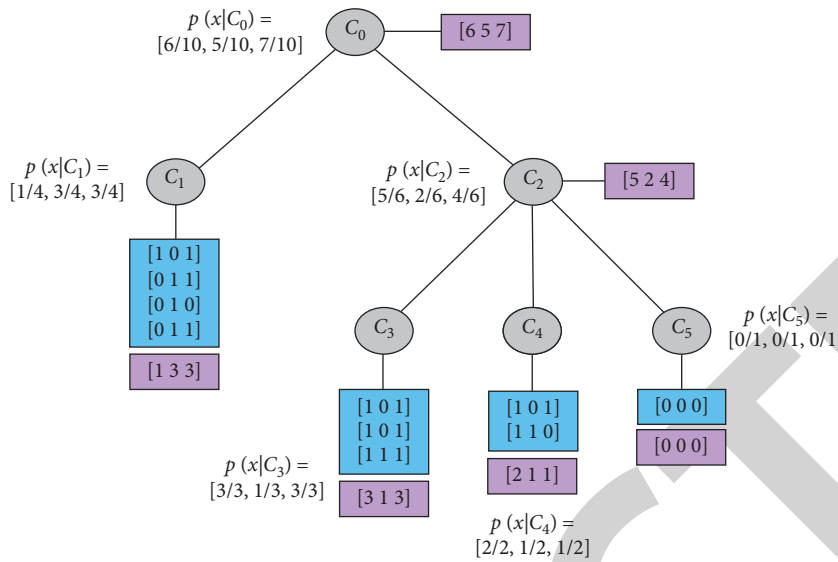

Figure 5: COBWEB clustering.

Table 2: College sports participants' characteristics.

	Age (yrs.)	BMI (kg/m^2)	BSA (m^2)
Min	14.00	17.65	1.440
1st qu	19.00	21.23	1.760
Median	32.00	22.65	1.840
Mean	32.00	22.76	1.832
3rd qu	41.00	24.24	1.910
Max	61.00	33.26	2.130

Table 2 shows the sampled sports students. Their ages ranged from 14 to 61 years, BMI ranges were 17.65 to 33.26 kg/m^2, and BSA ranged from 1.440 to 2.130 m^2.

```
In [2]: import numpy as np
        import pandas as pd

In [3]: import matplotlib.pyplot as plt
        import seaborn as sns
        %matplotlib inline

In [4]: from sklearn.datasets import make_blobs

In [5]: data = make_blobs(n_samples=700, n_features=10, centers=4)

In [6]: data[0].shape # features
Out[6]: (700, 10)

In [7]: data[1].shape # clusters column
Out[7]: (700,)
```

Figure 6: Python code with imports and data characteristics.

2.4.2. DBSCAN. The epsilon neighborhood of point p of our dataset D was defined as per equation (1), representing the core regions as $|N(p)|$>minimum number of points:

$$N(p) = \{q \in D | \text{dist}(p, q) \leq \varepsilon\}. \quad (2)$$

Core points are defined as such because they lie on the interior of the cluster. Border points are the ones that lie in the neighborhood of another core point. Noise, on the other hand, represents neighbor border or core points [20]. We employed the Sklearn python library for DBSCAN clustering. After importing the CSV data via the Panda library into a data frame, we dropped all the null values in all the columns and transformed the irrelevant data points. With 3 clusters, the code in Figure 9 was generated using the Matplotlib python library.

Figure 9 shows DBSCAN with 3 clusters. The picture was generated via the Matplotlib Python library.

3. Methodology Results

We observed that the k-means and EM clustering algorithms beat the others described in the previous sections when we examined the clustering methods utilizing accurate data. Both of these methods are capable of clustering data and

```
In [11]: sse_error = []
         for n_clusters in range(1,11):
             kmeans = KMeans(n_clusters)
             kmeans.fit(df)
             sse_error.append(kmeans.inertia_)
```

Figure 7: k-means clustering.

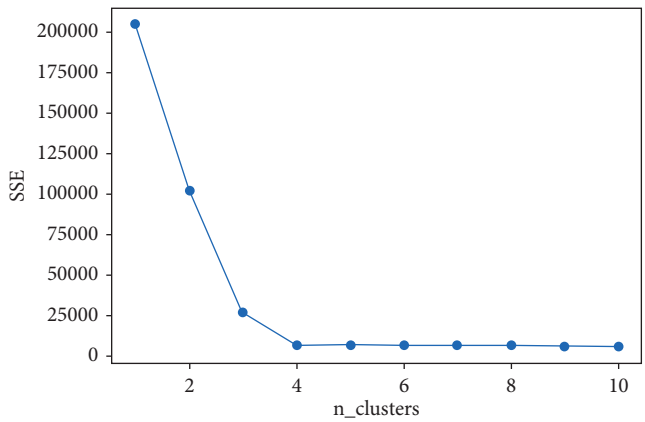

Figure 8: Elbow process graph representation with Matplotlib.

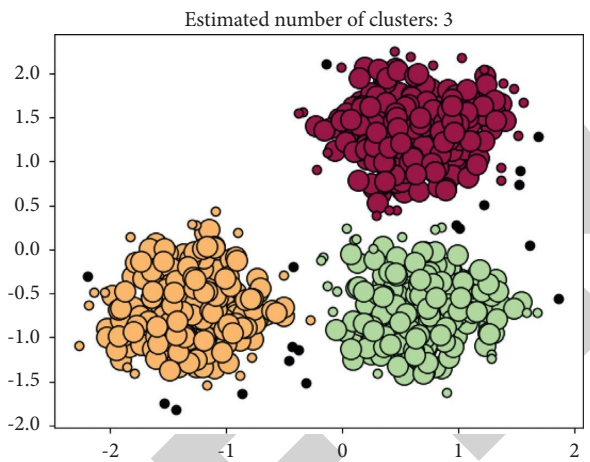

Figure 9: DBSCAN pictorial representation.

defining categories. The COBWEB and DBSCAN methods did not provide the same results. We used a hierarchical technique based on the DTW distance for dynamic data and compared the results to a hierarchy based on the Euclidean distance [21]. The Rand index [6] is used to compare two hierarchical clustering methods.

Consequently, when the number of clusters is large, the DTW hierarchical method is essentially similar to the previous technique. If the two tests have different durations, DTW is the better option for determining the distance. Maximum similarity grouping is based on work. The "clustering in cascade" method is employed in the third phase of the load. In this instance, k-means was used to examine the factual data and EM, while hierarchical clustering investigated the DTW distance's dynamic data.

We cluster similarity based on the domain expert's score to put the findings to the test. The data from domain experts are split into two files. The first file's Rand index values are all greater than 0.74. The values are about 0.8, mainly for the cascade clustering. It shows that our findings and the vote of domain experts [1] are very similar. The techniques used are effective. When the number of clusters is the same, k-means with hierarchical clustering beats EM with hierarchical clustering. In addition, the hierarchical clustering result is better. The rand index value in the second file is adequate but not as good as before the merge.

4. Conclusion

The use of data mining ideas and methods in sports has yet to develop to its full potential. Most sports organizations have just started to uncover facts and knowledge buried in their data in many aspects.

We examined an overview of data mining throughout this article and reviewed some of the methods utilized in data mining. We employed the clustering method, which uses an unsupervised learning process for college sports data mining. COBWEB, DBSCAN, k-means, and hierarchical methods are some of the approaches used for clustering. Based on the findings collected, a comparison was made. Because of their consistent results, we suggested k-means and EM algorithms, unlike DBSCAN and COBWEB methods.

The research's subsequent development is to automate classification devices to allocate a new test to the correct cluster to determine maximum workload and final score. Data mining is ubiquitous, and sports at higher learning institutions have long been done. It is thus of most tremendous significance to improve analytical processes via data mining for learning institutions and the sports fraternity.

In a word, the data mining technology is a helpful instrument that has been used in many areas and has been successful. Of course, data mining alone is not all-powerful, and its use cannot be isolated from the realistic backdrop. The technology will only have real life by adopting the people-oriented concept. However, the psychological system of sports includes a great deal of valuable knowledge to be found.

Data Availability

The data underlying the results presented in the study are included within the manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

Authors' Contributions

The author has read the manuscript and approved for submission.

References

[1] J. Han, M. Kamber, and D. Mining, "Concepts and techniques," *Morgan Kaufmann*, vol. 340, pp. 94104–103205, 2006.

[2] Y. Aref, K. Cemal, Y. Asef, and S. Amir, "Automatic fuzzy-DBSCAN algorithm for morphological and overlapping datasets," *Journal of Systems Engineering and Electronics*, vol. 31, no. 6, pp. 1245–1253, 2020.

[3] M. Cheffena and M. Mohamed, "The application of lognormal mixture shadowing model for B2B channels," *IEEE sensors letters*, vol. 2, no. 3, pp. 1–4, 2018.

[4] H. Chen, P. Buntin Rinde, L. She, S. Sutjahjo, C. Sommer, and D. Neely, "Expert prediction, symbolic learning, and neural networks: an experiment in greyhound racing," *IEEE Expert*, vol. 9, 1994.

[5] W. Deze, "Application of comprehensive data mining technology in colleges and universities," in *Proceedings of the 2nd International Conference on Big Data Research*, pp. 86–89, Weihai, China, 2018, October.

[6] M. Marconi, M. Germani, M. Mandolini, and C. Favi, "Applying data mining technique to disassembly sequence planning: a method to assess effective disassembly time of industrial products," *International Journal of Production Research*, vol. 57, no. 2, pp. 599–623, 2019.

[7] J. E. Chacón, "A close-up comparison of the misclassification error distance and the adjusted Rand index for external clustering evaluation," *British Journal of Mathematical and Statistical Psychology*, vol. 74, no. 2, pp. 203–231, 2021.

[8] A. J. Palm, "Using business intelligence to analyze sports associations' financial data," Masters thesis, University of Twente, Enschede, The Netherlands, 2021.

[9] I. Popovych, V. Zavatskyi, O. Tsiuniak et al., "Research on the types of pre-game expectations in the athletes of sports games," *Journal of Physical Education and Sport*, vol. 20, 2020.

[10] A. Rad, B. Naderi, and M. Soltani, "Clustering and ranking university majors using data mining and AHP algorithms: a case study in Iran," *Expert Systems with Applications*, vol. 38, no. 1, pp. 755–763, 2011.

[11] F. Ros and S. Guillaume, "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," *Expert Systems with Applications*, vol. 128, pp. 96–108, 2019.

[12] Z. Shelly, R. Burch, W. Tian, L. Strawderman, A. Piroli, and C. Bichey, "Using K-means clustering to create training groups for elite American football student-athletes based on game demands," *International Journal of Kinesiology & Sports Science*, vol. 8, no. 2, Article ID 47, 2020.

[13] S. Wang, "Research on weekly load characteristics and physical effect of physical education training based on data mining technology," *Converter*, vol. 2021, no. 6, pp. 646–653, 2021.

[14] Z. Yin and W. Cui, "Outlier data mining model for sports data analysis," *Journal of Intelligent and Fuzzy Systems*, vol. 40, pp. 1–10, 2021.

[15] X. Zhang and P. Luo, "Analysis of psychological education factors based on computer software and hardware collaboration and data mining," *Microprocessors and Microsystems*, vol. 81, Article ID 103744, 2020.

[16] P. Zhou, "Selection of cross-border E-commerce import model based on intelligent data analytics AHP algorithm," *Mobile Information Systems*, vol. 2021, Article ID 1351178, 8 pages, 2021.

[17] D. Godoy-Izquierdo, I. Díaz Ceballos, M. J. Ramírez Molina, E. Navarrón Vallejo, and J. Dosil Díaz, "Risk for eating disorders in "high"-and "low"-risk sports and football (soccer): a profile analysis with clustering techniques," *Revista de Psicologia del Deporte*, vol. 28, no. 2, pp. 0117–0126, 2019.

[18] K. Lee, H. W. Kim, C. Moon, and Y. Nam, "Analysis of vocal disorders using cobweb clustering," in *Proceedings of the 2019 International Conference on Artificial Intelligence in Information and Communication (ICAC)*, pp. 120–123, Okinawa, Japan, 2019 February.

[19] H. Li, J. Liu, Z. Yang, R. W. Liu, K. Wu, and Y. Wan, "Adaptively constrained dynamic time warping for time series classification and clustering," *Information Sciences*, vol. 534, pp. 97–116, 2020.

[20] L. S. Luteberget, B. R. Holme, and M. Spencer, "Reliability of wearable inertial measurement units to measure physical activity in team handball," *International Journal of Sports Physiology and Performance*, vol. 13, no. 4, pp. 467–473, 2018.

[21] N. Ma, "Neural network data mining application in health education and intervention effect of health risk behaviors among college students," *Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. 23, no. 6, pp. 226–237, 2016.