**International Journal of Intelligent Information Technologies**
Volume 20 • Issue 1

# The Promotion of Women's Leisure Sports Behavior Based on Improved Decision Tree Algorithm

Huaping Luo, Hunan College of Chemical Technology, China*

## ABSTRACT

In women's daily leisure choices, sports is an important content that cannot be ignored. In this context, this paper studies the promotion of women's leisure sports behavior based on improved decision tree algorithm. Based on the simple analysis of the research progress of leisure sports and decision tree algorithm, a female leisure sports behavior model based on decision tree is constructed. Based on the decision tree algorithm, the calculation method of information gain rate is optimized to avoid logarithmic operation, and the continuous attributes are discretized. Simulation results show that in terms of classification accuracy, the improved decision tree algorithm is significantly higher than the classical decision tree algorithm, and can significantly shorten the running time, which has high application value in the realization of accurate classification analysis of female leisure sports behavior.

## KEYWORDS

Decision Tree Algorithm, Female Leisure Sports Behavior, Information Gain Rate, Taylor Formula

## INTRODUCTION

With the improvement of national physical quality and economic development, sports have become an indispensable part of life, and leisure sports have gradually become an important part of people's daily leisure lives (Yu, 2020). Women living in the new era have a great demand for leisure sports and have also spent a lot of energy and financial resources on leisure sports (Zheng, 2018). At present, the research on sports behavior is mostly carried out in cities, considering things such as sports consumption, sports choice, sporting goods, and stadiums, but it is rarely discussed from the perspective of women (Strain et al., 2020).

This paper studies the promotion of women's leisure sports behavior based on an improved decision tree algorithm, which is mainly divided into four sections (Ahn & Chon, 2018). The first section briefly introduces the research background of leisure sports behavior and the arrangement of this study. The second section introduces the research status of behavior at home and abroad and the application and improvement of the decision tree algorithm. This section also summarizes

DOI: 10.4018/IJIIT.334709 *Corresponding Author

the shortcomings of current research. The third section constructs an analysis model of women's leisure sports behavior based on a decision tree model, improves the decision tree algorithm and the calculation method of information acquisition, and discretizes the continuous attribute problem. In the fourth section, the improved decision tree model constructed in this paper is simulated and analyzed to test the accuracy and running time of the algorithm. Compared with the classical decision tree algorithm, the experimental results show that the improved decision tree algorithm proposed in this paper has advantages in classification accuracy and running time, and has good application value in the analysis of female leisure sports behavior.

The innovation of this paper is the improvement strategy of decision tree analysis using the Taylor formula and McLaughlin expansion formula to improve the information gain rate of the C4.5 approximation algorithm, simplify the algorithm, improve the accuracy of the algorithm, and avoid logarithmic operation. In addition, aiming at the problem of discretization of continuous attributes, the discretization method based on the chi-square value is used to obtain the alternative optimal splitting breakpoint value to ensure classification accuracy.

## LITERATURE REVIEW

There is much research on behavior analysis at this stage. The early research is mainly qualitative research. With the development of big data technology, data mining technology is gradually applied. By mining the action data extracted from the log files of different subsystems in the corresponding field, different types of user behaviors in different modules and subsystems in the intelligent manufacturing environment are discovered and determined (Shang et al., 2020). In the network behavior analysis, Tao et al. (2019) analyzed the microblog behavior through machine learning and cloud computing technology and proposed a microblog recommendation algorithm based on statistical features (Tao et al., 2019). The feature data mining is carried out through a cloud computing big data method, which is suitable for online mining microblog behavior (Yuan, 2020). In the analysis of consumer behavior, Dixit et al. (2021) established a multilevel hesitation mining model. The multi-stage hesitation model improves the threshold and confidence threshold in design and the minimum support threshold covers attraction and hesitation and uses prior attributes to generate a mining model in a step-by-step. In their research, Dixit et al., in the analysis of behavior characteristics, sorted out the characteristics of customer data and the loss of existing customers and used a fuzzy decision tree algorithm to establish a decision tree. The accuracy of the iris data set reached 97.8%, achieving the highest accuracy (Hu et al., 2020). Han et al. also optimized and improved the decision tree algorithm in this study, introduced temporal feature selection, combined it with the decision tree algorithm, used ant miner fuzzy decision tree classifier to extract intelligent fuzzy rules from weighted temporal capabilities, and then used fuzzy rule extractor to reduce the diversity of functions in the extracted rules (Han et al., 2021). Choi et al. designed a new decision tree structure. After analyzing the experimental motion pattern data using the original value of a single inertial measurement unit in a 200 ms time window, they identified nine common motion patterns hierarchically, and used artificial bee colony algorithm to search the initial weight and threshold globally (Choi et al., 2018).

To sum up, we can see that there is much research on behavior analysis and decision tree algorithms at home and abroad. In addition to qualitative analysis, data mining technology is widely used in behavior analysis, and the decision tree algorithm also has many improvements based on classical algorithms. However, these decision tree algorithms are mostly used to analyze the behavior of a certain type of data, and their adaptability is not strong. In addition, in terms of behavior analysis, leisure sports behavior analysis mostly belongs to qualitative analysis. Quantitative research is insufficient and rarely discusses women's vision, mostly from the perspective of mass consumption behavior, so the research is not universal. Therefore, it is of great practical significance to carry out research on women's leisure sports behavior based on an improved decision tree

algorithm. This study focuses on promoting women's leisure sports behavior by improving decision tree algorithms. We have established a model for analyzing women's leisure sports behavior and optimized the improved algorithm to address the shortcomings. Specifically, it includes improving the calculation method of information gain rate and using chi-square values based on continuous attributes to merge adjacent intervals to reduce candidate breakpoints. This improved algorithm improves classification accuracy.

## METHODOLOGY

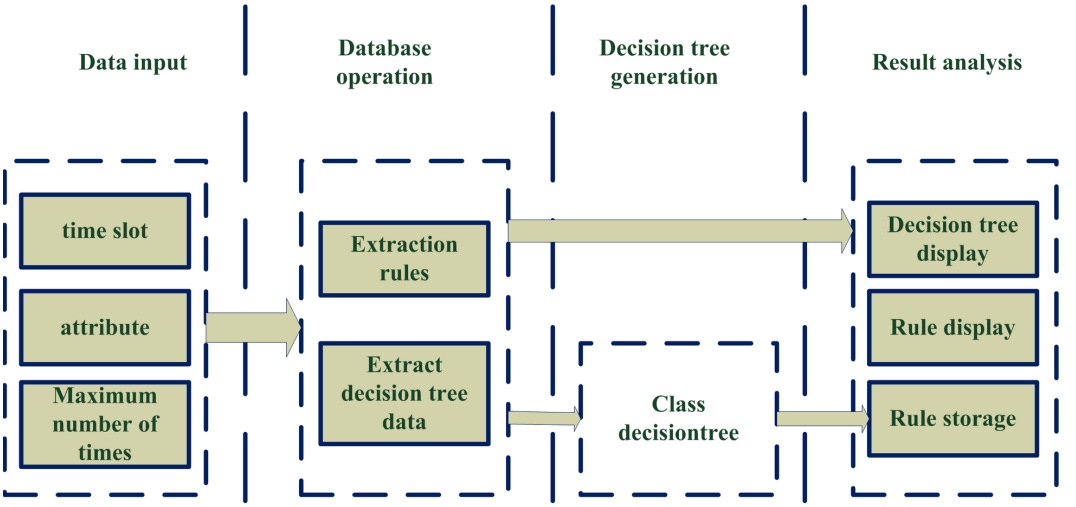
### Modeling of Female Leisure Sports Behavior

The modeling of women's leisure sports behavior and the analysis of data mining algorithms can timely grasp and predict women's leisure sports behavior habits (Hou, 2021). At present, the development of big data technology records a lot of women's leisure sports data, covering women's basic information, product and service information, consumption information, etc (Xie et al., 2021). Analyzing women's leisure sports behavior knowledge from these massive data can group women, observe behavior habits, and tap potential development opportunities. Women's leisure sports behavior analysis should be able to achieve clustering combined with behavior analysis. In principle, the smaller the group, the better, form a target model, analyze women's leisure sports preferences, and provide more targeted services for women according to women's leisure sports preferences.

A decision tree algorithm is a basic classification and regression method that analyzes the characteristics of data and constructs a tree structure with features as nodes and feature values as branches, thereby achieving data classification and prediction (Mao & Zhang, 2021). The generation process of a decision tree can be achieved recursively by selecting the best features for node splitting until the stopping condition is met (such as when the number of node samples is less than the threshold). At each node, the algorithm selects a feature and divides the data into different subsets based on its value until a certain stopping condition is met. The commonly used feature selection criteria include information gain, Gini index, and mean square error, which have various applications in classification and regression problems (Pham et al., 2021). Information entropy is used to measure the uncertainty of data, while information gain refers to the degree to which information entropy decreases after data is divided by features, in order to select the optimal partitioning feature (Lou et al., 2021). The key to decision tree algorithms is how to select the optimal features for node splitting. Common generation algorithms include iterative dichotomiser 3 (ID3), C4.5, and classification and regression tree. To prevent overfitting of the decision tree, pruning operations are required. Pre-pruning determines whether to split nodes before splitting, while post-pruning prunes after generating a complete tree. Pruning operations can improve the generalization ability of the model. Decision tree algorithms have wide applications in fields such as data mining, medical diagnosis, and financial risk control. Their simple and intuitive characteristics make them very popular in highly explanatory scenarios. In this paper, a decision tree algorithm is used to analyze women's leisure sports, which can be roughly divided into input, operation, generating decision tree, and result processing, as shown in Figure 1.

User input mainly includes time period, attribute information, and maximum number of times. Users can input time information directly. Attribute information records the attribute information of women's leisure sports behavior. Maximum number of times receives input numbers using a textbox. This information is transferred to the Data Access Layer through the Business Logic Layer, and the maximum number of times is selected to be transferred to the decision tree generation class. Among many big data mining algorithms, the decision tree algorithm is a classic type of classification algorithm. This algorithm has a simple structure and is only a tree structure result, which is very clear (Ducange et al., 2021). The decision tree algorithm can be constructed in a very short time and has high efficiency, especially when dealing with big data. The classification results of the decision tree algorithm have high accuracy, high scalability, and can be applied to

**Figure 1. Female leisure sports behavior analysis model**




a variety of databases (Ducange et al., 2021). Extract the existing rules and decision tree data in the operation of the database. If the decision tree has been generated, directly extract the rules from the database. If it is the first generation, the data of the generated decision tree is extracted. The improved decision tree algorithm is used to generate the decision tree. In the current decision tree algorithm, the C4.5 algorithm is widely used, but it can only perform the optimal dichotomy once, so it needs to be improved. In the modeling, we use the results of the web system and Oracle database to mine the female leisure sports behavior.

## Improvement of Decision Tree Algorithm

The ID3 algorithm is one of the earliest of the many decision tree algorithms. In the process of constructing a decision tree, non-class attributes are represented by nodes of the decision tree, and non-type possible values are represented by edges when dividing and selecting attributes, according to the decline speed of information entropy, when selecting test attributes, according to the path to the current node. ID3 is a classical algorithm, which does not analyze the conditional attribute of the maximum information gain. For non-terminal successor nodes, the same process is used to segment training samples. ID3 algorithm is a decision tree algorithm with high practical value. Its non-basic theory is clear and easy to understand, but it also has its limitations.

ID3 algorithm has the problem of multi-value bias. When selecting decision attributes, it will give priority to the conditional attributes with more attribute values (Cherfi et al., 2018). However, in many cases, the attribute value with more values is not the optimal attribute. When constructing the decision tree, the nodes have only one attribute. There is not much correlation between the attributes when using the univariate algorithm. Therefore, although they are connected to the decision tree, they are still scattered attributes. The ID3 algorithm is easily affected by noise data. If the noise is not removed, the wrong eigenvalue may be selected. Compared with other algorithms, the ID3 decision tree algorithm has a clear theory, but it is difficult to calculate and needs a long training time and space.

Given the shortcomings of the ID3 algorithm, many studies have improved the algorithm. The C4.5 algorithm is based on the ID3 algorithm (Li et al., 2020). Compared with the ID3 algorithm, the information gain rate is used as the standard when selecting attributes, and the continuity value can be processed when processing data, which increases the data range. If the processing

is not complete, it can also be processed to improve the adaptability of the data. This algorithm needs continuous construction operation in the framework model, so it can avoid the problem of imbalance (Guerrero et al., 2020). The classification rules in the decision tree are represented by if-then. Compared with the ID3 algorithm, it can handle the numerical cases of continuous attributes and vacant attributes, so the efficiency is improved. Specifically, assuming the training sample set adopts representation when constructing the decision tree of the training sample set, select the attribute with the largest information gain rate as the splitting node, so that the sample set can be divided into multiple subsets (Chang et al., 2023). If the tuple categories contained in the subset are consistent, this node can be used as a child node of the decision tree. If this condition is not met, the same method is used to generate the tree until all the elements contained in the subset belong to the same category.

Assuming that the training set is represented by $S$ and divided into $m$ classes, the calculation formula based on information entropy is:

$$Info(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Where $p_i$ represents probability and $Info$ represents category information entropy. If the sample set is divided into multiple subsets after attribute $A$ is divided into training set $S$, assuming that attribute $A$ has different values, a total of $k$, the calculation formula of information entropy of $A$ divided into subsets is as follows:

$$Info_A(S) = \sum_{j=1}^{k} \frac{s_{1j} + s_{2j} + \ldots + s_{mj}}{s} \times Info(s_j) \tag{2}$$

Where $Info_A$ represents conditional information entropy, $p_{ij}$ represents sample probability, and $p_{ij} = \frac{s_{ij}}{s_j}$. The information gain calculation formula of division attribute $A$ is expressed as:

$$Gain(A,S) = Info(S) - Info_A(S) \tag{3}$$

Attribute $A$ has different values, and the sample set is $A$ divided into different subsets. Assuming that some samples have the value $a_j$ on the attribute $A$, the samples are segmented with $A$ as the benchmark value to obtain the split information entropy. The formula is expressed as:

$$Info(A) = -\sum_{j=1}^{k} p_j \log_2(p_j) \tag{4}$$

The formula of information gain rate for dividing attribute $A$ is expressed as:

$$Gain - Radio(A) = \frac{Gain(A,S)}{Info(A)} \tag{5}$$

In the application of the algorithm, the training sample set is input, the decision attributes are output, and the decision tree is established with the training sample set as the root node (Aghaabbasi et al., 2021). If all samples belong to the same category, record them as leaf nodes and mark the category. Calculate the gain rate of attribute information in the set, select the largest as the segmentation attribute of the node, determine the subset of training samples according to this value, and generate the corresponding branches. Make full use of this step to generate new branches until there are no sub-nodes to be divided, as shown in Figure 2.

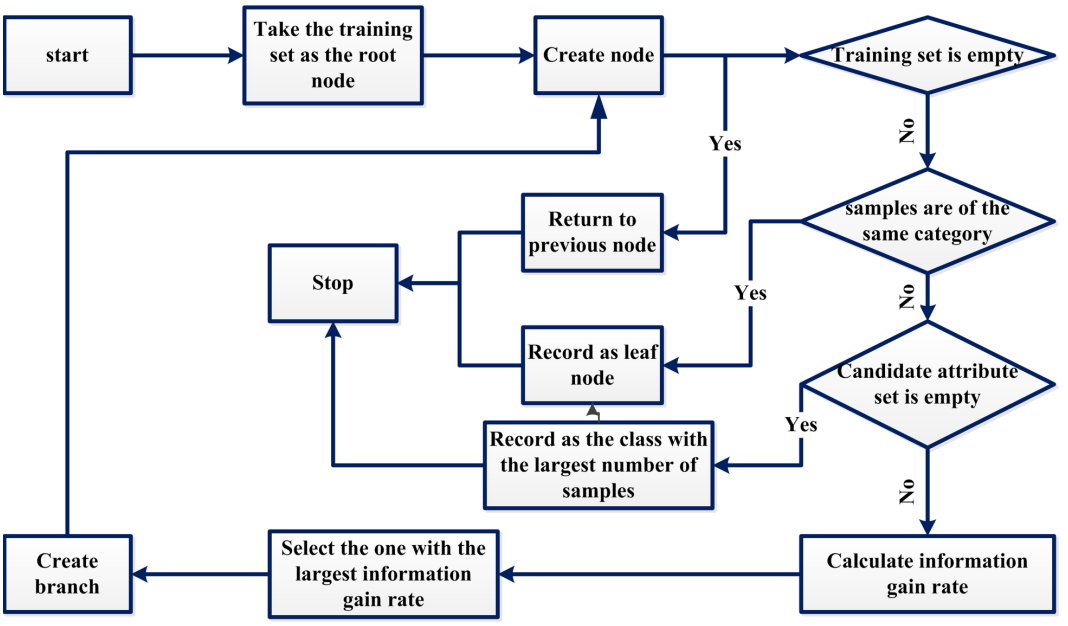
The decision tree algorithm is mainly aimed at optimizing the time continuity attribute. The information gain rate is used in the evaluation of classification ability, and the method with a large information gain rate is used in the construction of direct points of the decision tree (Bôto et al., 2022). However, the partition information will gradually tend to zero, and the ratio is not stable enough, so constraints need to be added. The information gain rate required to be tested is very large. Assuming that the conditional attribute set is represented by $C$, the continuous attribute set is represented by $C1$, the discrete attribute set is represented by $C2$, and $a$ is the attribute of the discrete attribute set, the $a$ is sorted. The information gain rate defined for the continuous attribute $a$ can be expressed as:

$$GainRatio(a,cut) = \frac{Gain(a,cut)}{SplitInfo_a(CK,cut)} \tag{6}$$

Where $cut \in cuts(a)$, if $a(t) < cut$, then the $a(t)$ threshold value is 0, and $a(t)$ is 1 in other cases. Homogenize attributes according to breakpoints. The information gain rate of continuous attribute about breakpoint $a$ can be expressed as:

$$opt\_cut(a) = \arg\max\{GainRatio(a,cut)\}, \forall cut \in cuts(a) \tag{7}$$

**Figure 2. Decision tree algorithm flow**



When selecting the node splitting attribute, the algorithm needs to calculate the gain rate of each information and select the largest data as the column attribute, which will involve multiple logarithmic operations and call database functions, which increases the difficulty of calculation (Dong et al., 2019). In the discretization of continuous attributes, it is necessary to calculate the information gain rate of all segmentation points, which leads to the extension of the running time of the algorithm. For classification attributes, the attribute value is related to the number of node branches, but there are many empty branches in the actual calculation, which has no impact on the classification. These branches may cause overfitting problems and affect the application effect of the decision tree for C4.5 problems in the algorithm.

Taylor series is expressed by function and belongs to the infinite term. The infinite term is obtained by calculating the reciprocal of the function. After removing the Lagrange remainder, the approximate formula can be expressed as:

$$f(x) \approx f(0) + f'(0)x + \frac{f'(0)x^2}{2!} + \ldots + \frac{f'(0)x^n}{n!} \tag{8}$$

Since the reciprocal of Ln (x) is meaningless when x is zero, and the information gain rate is 0 ~ 1 in the value range of the calculation formula(9), it can be considered to improve the formula by using McLaughlin consensus:

$$\ln(x+1) \approx x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + \ldots + (-1)^n \frac{1}{n}x^n \tag{9}$$

Further converted into:

$$\ln(x) \approx (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 + \ldots + (-1)^n \frac{1}{n}(x-1)^n \tag{10}$$

When x ranges from 0 to 1, the information gain rate is:

$$\ln(x) \approx (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 \tag{11}$$

After the above approximate simplification, the logarithmic operation has been eliminated and the complex calculation process has been avoided. The category information is converted to:

$$Info'(S) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s} \tag{12}$$

Similarly, conditional information entropy and split information entropy can be obtained, and the formula is:

$$Info_A(S) = -\frac{1}{\ln 2 \times s} \sum_{j=1}^{k} \sum_{i=1}^{m} [\frac{s_{ij}(s_{ij} - s_j)(11s_j^2 + 2s_{ij}^2 - 7s_{ij}s_j)}{6s_j^2}] \tag{13}$$

$$Info(A) = -\frac{1}{\ln 2 \times s}\sum_{j=1}^{k}[\frac{s_j(s_j - s)(11s^2 + 2s_j^2 - 7s_j s)}{6s^2} \tag{14}$$

The information gain rate obtained after conversion can be expressed as:

$$Gain - Ratio(A) = \frac{Info(S) - Info_A(S)}{Info(A)} \tag{15}$$

It can be seen from formula(15) that after a calculation, the information gain values on the category information are the same, which is because some information is omitted during the improvement, and the classification accuracy of the algorithm is improved. When improving, it is necessary to calculate the category conditional entropy to ensure the order of information gain rate and avoid affecting the classification accuracy. The traditional C4.5 algorithm needs a lot of calculation of logarithmic function when calling the function. After improvement, it only needs four simple clouds, which eliminates logarithmic calculation, so it can improve the operation efficiency of the algorithm.

The improved algorithm selects a higher information gain rate when dealing with continuous attribute values. To sort continuous attribute values, it is necessary to scientifically determine the number of discrete values. In this paper, the discretization method based on chi-square statistics is adopted, and the continuous attributes that need to be discretized are determined by using the relationship between continuous attributes and category attributes. Without changing the classification relationship, the chi-square value of each adjacent interval is calculated, and the adjacent intervals are combined to obtain the optimal breaking point. In the application, first, judge whether the attributes in the set are discrete. If not, calculate the information gain rate for the discrete attributes. If not, sort according to the order from large to small, form the initialization interval, obtain the chi-square value, merge the intervals with a small chi-square value, obtain the candidate splitting breakpoints, calculate the information gain rate, and determine the splitting attributes and breakpoints.

## RESULT ANALYSIS AND DISCUSSION

### Performance Test of Improved Decision Tree Algorithm

The data set in UCI (University of California, Irvine) data is used for performance tests, and there is only one category attribute of all data sets. To ensure the adaptability of the experimental results, the examples in all data sets are divided into two groups, with 50% as the training sample set and other data as the test sample set. Statistical analysis was conducted on the performance differences between improved decision tree algorithms, ID3 algorithms, cart algorithms, C4.5 algorithms, and CART algorithms.

The classification accuracy of the improved decision tree algorithm is compared and analyzed. The test results are shown in Figure 3. From the data in Figure 3, we can see that C4.5 has some advantages over ID3. The classification accuracy of the improved decision tree algorithm is higher than that of the classical decision tree algorithm.

Compare and analyze the running time under different algorithms, and the measurement results are shown in Figure 4. It can be seen from the data in Figure 4 that the improved decision tree algorithm proposed in this paper does not prolong the running time, which shows that the improved decision tree algorithm can improve the accuracy on the basis of shortening the running time.

Considering that the accuracy of classification is affected by the change of parameters, it is necessary to optimize the parameters. In the model simulation, a relatively simple UCI library is used for analysis. The data were divided into two groups, 50% as the training sample set and other

Figure 3. Comparison of classification accuracy

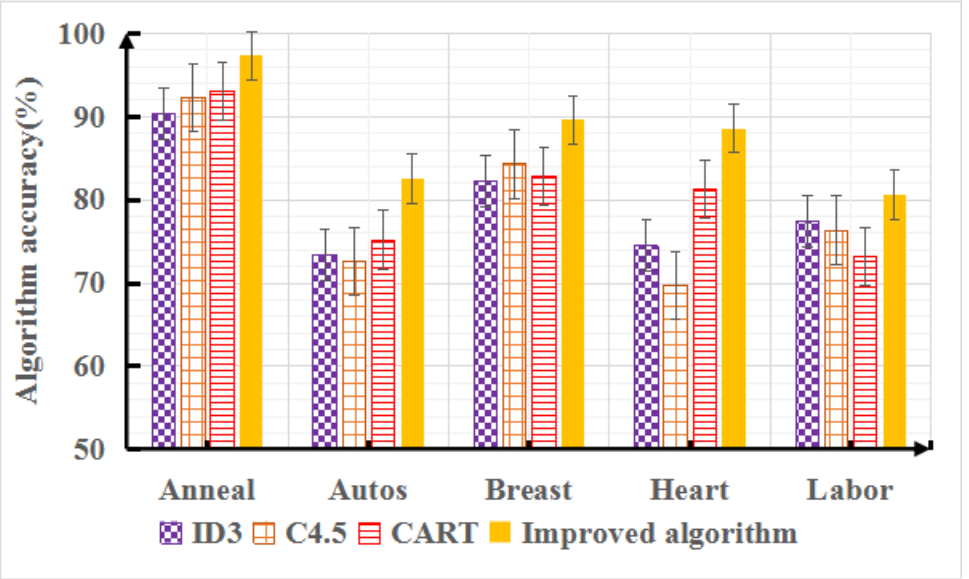
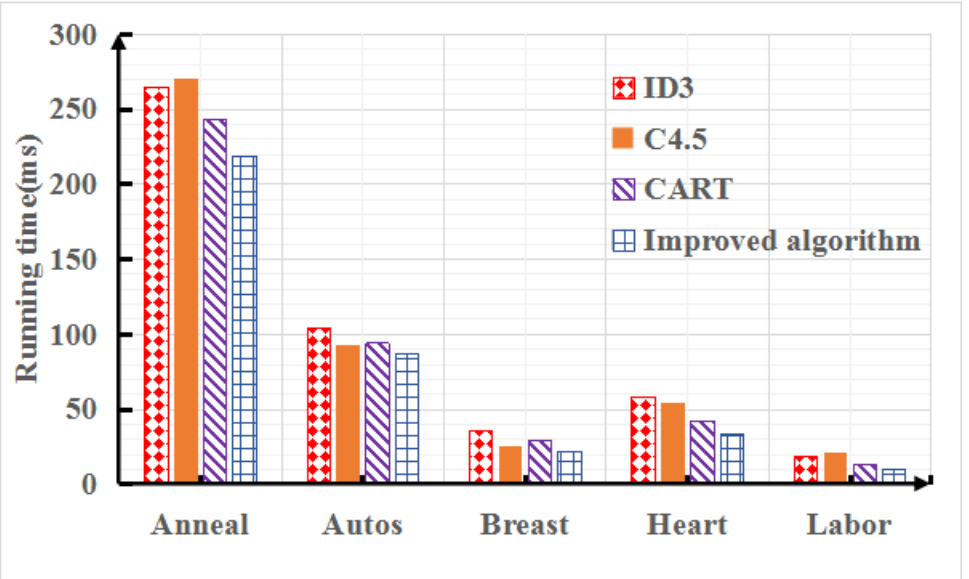



Figure 4. Comparison of algorithm running time




data as the test sample set. Change the size of the parameter. The measurements are shown in Figure 5. As can be seen from the data in Figure 5, there is a positive correlation between the early change in classification accuracy and the parameter value. But when the parameter is 12, the precision will not be improved significantly. Therefore, the parameter value is set to 12.

The improved decision tree algorithm adopts the chi-square discretization method in the discretization of continuous attributes. The algorithms before and after discretization are compared and analyzed to measure the classification accuracy and running time. The measurement results are

Figure 5. Classification accuracy under different parameters

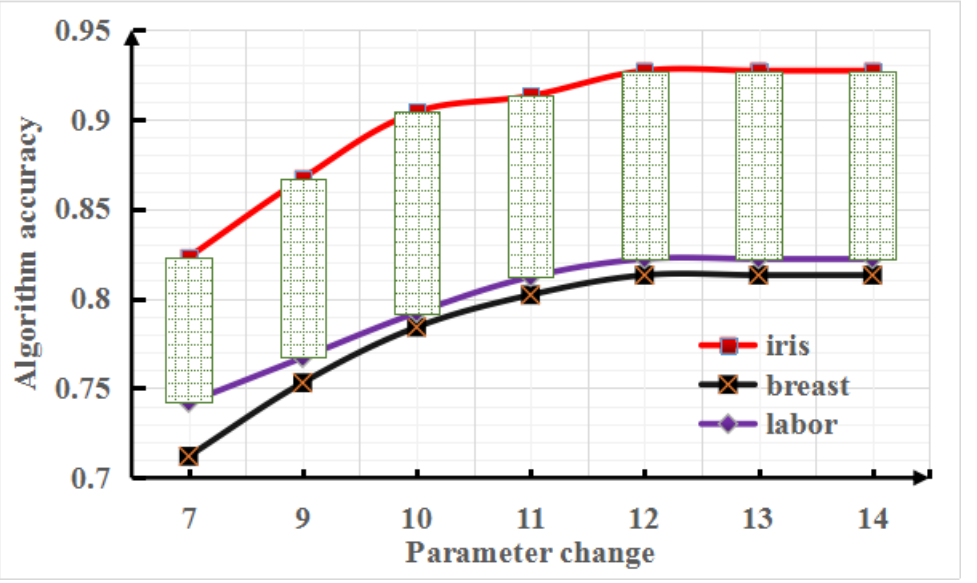
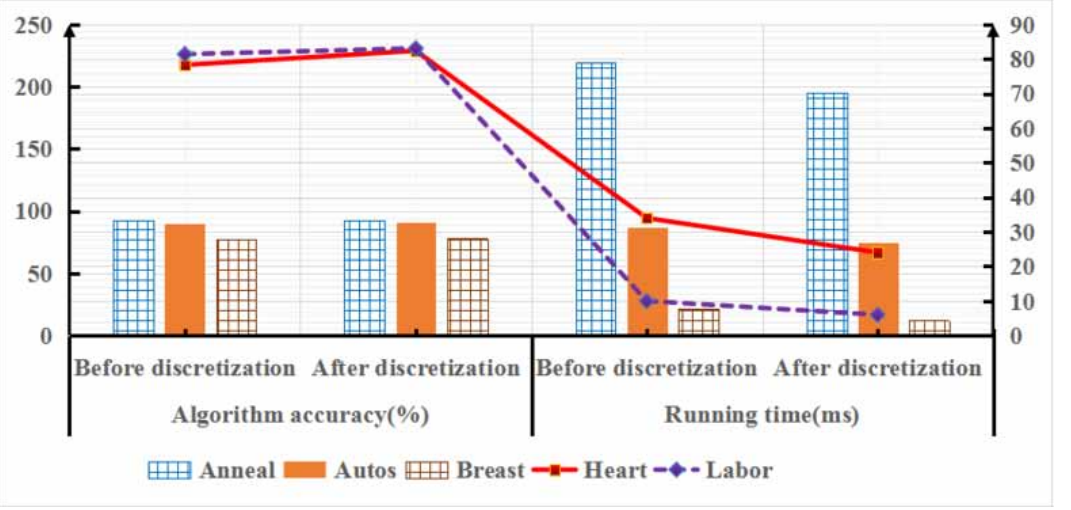



Figure 6. Effect of discretization on algorithm performance




shown in Figure 6. It can be seen from the data in Figure 6 that the running time of the improved algorithm after discretization processing is significantly shortened, and the classification accuracy does not decline, which proves the superiority of the algorithm.

## Simulation Analysis of Female Leisure Sports Behavior

Collect the data on women's leisure sports behavior. In addition to the basic information about women, the important basic information also includes the types of leisure sports, forms of participation, times of stadiums, and sports consumption. Before data analysis, it needs to be preprocessed to ensure privacy information and standardize data processing. The data information obtained from women's

leisure sports behavior is not standardized and cannot be analyzed directly. It needs to be transformed, including data clarity, integration, and transformation.

The original data is basically from various business databases, with many duplicates and missing information. The data needs to be clear. For the noise data with high expected values, the average value box method is used for smoothing. Remove the data irrelevant or redundant to women's leisure sports behavior, to avoid affecting the operation of the decision tree, or adopt the same clustering method to analyze the data. The data format is normalized, and the continuous variable data is discretized to meet the mining requirements of the decision tree algorithm. In data conversion processing, digital data occupies less storage space and faster calculation speed, so some characters are converted to digital types. Transform large capacity data into highly available data sets and adopt clustering method to standardize the data in preprocessing.

In the actual modeling of women's leisure sports behavior, data is the basis for establishment. The reference of different models can improve the classification accuracy to a certain extent and avoid overfitting. The boosting algorithm is introduced into the modeling, iterated for ten times, and 50% of the samples are selected for iterative calculation to obtain the decision tree model. Figure 7 is the application result of the model in the test data. From the data in Figure 7, we can see that the error rate decreases significantly, indicating that the boosting algorithm can improve the classification accuracy of the decision tree and avoid overfitting.

Taking female leisure behavior and sports behavior as an example, the decision tree model is established, and the results are shown in Figure 8. It can be seen from Figure 8 that women's leisure sports consumption has good consumption habits, and the consistency of the model training set is high, almost completely consistent, indicating that the female leisure sports behavior model designed in this paper based on the improved decision tree is suitable for behavior analysis.

## Application Analysis of Improving Decision Tree Algorithm

The improved decision tree algorithm has a significant impact on promoting women's leisure sports behavior. By researching and applying improved algorithms, we can more accurately predict

**Figure 7. Adding boosting algorithm error rate change**

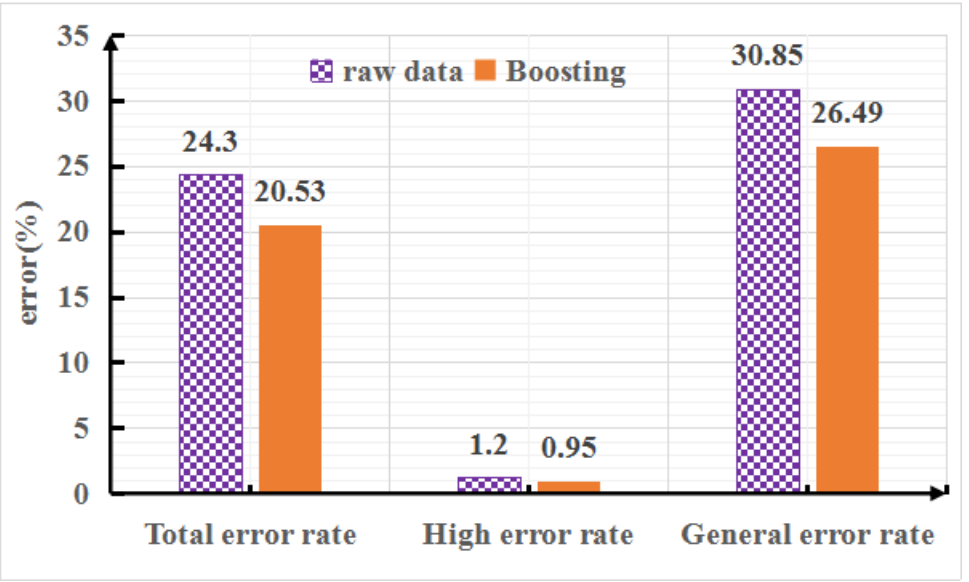
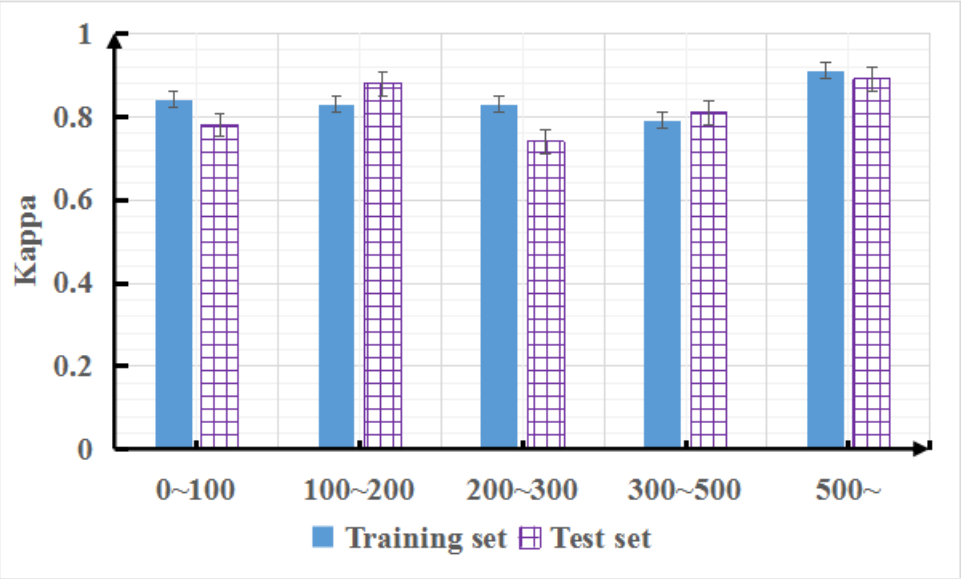


**Figure 8. Kappa parameters of decision tree model**




and analyze women's leisure sports behavior, which helps to better understand and promote the development of this field.

First, the improved decision tree algorithm has significant performance in behavior prediction. By optimizing the calculation method of information gain rate and merging adjacent intervals with chi-square values of continuous attributes, we can more accurately identify key factors affecting women's leisure sports behavior, thereby improving the accuracy and reliability of behavior prediction. This is of great significance for developing targeted promotion plans and resource allocation.

Second, the improved algorithm contributes to a more in-depth analysis of women's leisure sports behavior. By reducing candidate breakpoints and designing comparative experiments to analyze algorithm performance, we can have a more comprehensive understanding of the characteristics and patterns of women's leisure sports behavior, revealing potential factors and internal connections. This helps to provide more targeted advice and guidance to relevant decision-makers and promotes the positive development of women's leisure sports behavior.

Overall, the improved decision tree algorithm has improved the accuracy of predicting women's leisure sports behavior and deepened the breadth and depth of behavior analysis, providing strong support for promoting the development of women's leisure sports behavior.

However, in practical applications, in addition to calculating the information gain rate, some factors may affect the performance of decision tree algorithms. Targeted research will help improve the effectiveness of decision tree algorithms in practical applications and provide new solutions for addressing the challenges posed by the constantly increasing amount of data. Future research directions can include the following aspects:

1. Data quality and feature selection: Data quality is crucial for the performance of decision tree algorithms. In practical applications, problems such as data loss and noise interference may be encountered. Therefore, further research is needed on how to handle incomplete or noisy data to improve the robustness of the algorithm. At the same time, effective feature selection for large-scale datasets is also an important research direction to reduce dimensions and improve algorithm efficiency.

2. Unbalanced data processing: In practical scenarios, there may be category imbalance in female leisure sports behavior data, where the sample size of certain categories is much smaller than that of others. Future research can explore how to process imbalanced data through methods such as oversampling, undersampling, or ensemble learning to improve the model's recognition ability for minority class samples.
3. Large-scale data processing: With the continuous growth of data volume, how to effectively process large-scale data has become a challenge. Future research can explore technologies such as parallel computing and distributed algorithms to address the challenges posed by large-scale data on algorithm performance and computational efficiency.
4. Interpretability and interpretability: Decision tree algorithms typically have strong interpretability, but as models become more complex or face large-scale data, their interpretability may be affected. Therefore, future research can explore how to improve the interpretability of the model while maintaining its performance, so that decision-makers can better understand the prediction process and results of the model.

In addition, the decision tree algorithm is not limited to machine learning, it has also demonstrated strong application potential in many other fields. For example, in the field of medical diagnosis, decision tree algorithms can help doctors diagnose and predict diseases based on patients' symptoms and detection indicators. In the financial field, it is widely used in credit scoring, anti-fraud identification, and risk management. In marketing and customer relationship management, decision tree algorithms can perform market segmentation, customer classification, and personalized marketing recommendations based on customer characteristics and behavior patterns.

In addition, decision tree algorithms have demonstrated their strong application value in fields such as ecology, environmental science, and industrial manufacturing. This multi-domain application makes the decision tree algorithm a very practical and versatile algorithm tool, providing strong support for the development and progress of various industries. These cross-disciplinary developments not only promote the application of decision tree algorithms in various fields but also provide valuable experience and inspiration for their development in the field of machine learning, driving the continuous improvement and innovation of algorithms themselves, forming a virtuous cycle of development pattern.

## CONCLUSION

Among many data mining algorithms, the decision tree classification algorithm belongs to the top-down greedy algorithm. It selects the attribute with the best classification effect of each node and does not need more background knowledge. The sample data can directly generate the decision tree, which has many applications in behavior prediction and analysis. This paper studies the promotion of women's leisure sports behavior based on an improved decision tree algorithm. The analysis model of female leisure sports behavior is established to improve the shortcomings of the algorithm, optimize the calculation method of information gain rate, merge adjacent intervals based on chi-square value of continuous attributes, reduce candidate breakpoints, and design comparative experiments to analyze the performance of the algorithm. The results show that compared with the traditional decision tree algorithm, the improved decision tree algorithm improves the classification accuracy of the decision tree, avoids overfitting, and has higher consistency of the model training set. There are many factors affecting the performance of the decision tree algorithm. In addition to the calculation of information gain rate proposed in this paper, there are also conditional attribute relationships, front and back pruning, etc. Moreover, the amount of data in the application is increasing, and the newly generated data may affect the obtained results, which need to be further studied.

## AUTHOR NOTE

The figures used to support the findings of this study are included in the article.

The authors of this publication declare there are no competing interests.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author(s) of the article.

The authors would like to show sincere thanks to those techniques who have contributed to this research.

## REFERENCES

Aghaabbasi, M., Shah, M. Z., & Zainol, R. (2021). Investigating the use of active transportation modes among university employees through an advanced decision tree algorithm. *Civil and Sustainable Urban Engineering*, *1*(1), 26–49. doi:10.53623/csue.v1i1.28

Ahn, B. W., & Chon, T. J. (2018). Verification of relationships of serious leisure, leisure facilitator and psychological happiness among leisure sports participants in South Korea. *Asia Life Sciences, 15*(2), 1331–1339. http://scholarworks.bwise.kr/ssu/handle/2018.sw.ssu/34385

Bôto, J. M., Marreiros, A., Diogo, P., Pinto, E., & Mateus, M. P. (2022). Health behaviours as predictors of the Mediterranean diet adherence: A decision tree approach. *Public Health Nutrition*, *25*(7), 1864–1876. doi:10.1017/S1368980021003293 PMID:34369348

Chang, Y., Warren, C., & Katz, M. (2023). Determinants of subscription renewal behavior in sport spectatorship services: A CHAID decision tree modeling approach. *Sport Marketing Quarterly*, *32*(2), 124–136. doi:10.32731/SMQ.322.062023.03

Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very fast C4.5 decision tree algorithm. *Applied Artificial Intelligence*, *32*(2), 119–137. doi:10.1080/08839514.2018.1447479

Choi, J., Song, E., & Lee, S. (2018). L-Tree: A local-area-learning-based tree induction algorithm for image classification. *Sensors (Basel)*, *18*(1), 306. doi:10.3390/s18010306 PMID:29361699

Dixit, A., Tiwari, A., & Gupta, R. K. (2021). A model for trend analysis in the online shopping scenario using multilevel hesitation pattern mining. *Mathematical Problems in Engineering*, *2021*, 1–11. doi:10.1155/2021/2828262

Dong, W., Cao, X., Wu, X., & Dong, Y. (2019). Examining pedestrian satisfaction in gated and open communities: An integration of gradient boosting decision trees and impact-asymmetry analysis. *Landscape and Urban Planning*, *185*, 246–257. doi:10.1016/j.landurbplan.2019.02.012

Ducange, P., Marcelloni, F., & Pecori, R. (2021). Fuzzy Hoeffding decision tree for data stream classification. *International Journal of Computational Intelligence Systems*, *14*(1), 946–964. doi:10.2991/ijcis.d.210212.001

Guerrero, M. D., Vanderloo, L. M., Rhodes, R. E., Faulkner, G., Moore, S. A., & Tremblay, M. S. (2020). Canadian children's and youth's adherence to the 24-h movement guidelines during the COVID-19 pandemic: A decision tree analysis. *Journal of Sport and Health Science*, *9*(4), 313–321. doi:10.1016/j.jshs.2020.06.005 PMID:32525098

Han, Y., Liu, C., Yan, L., & Ren, L. (2021). Design of decision tree structure with improved BPNN nodes for high-accuracy locomotion mode recognition using a single IMU. *Sensors (Basel)*, *21*(2), 526. doi:10.3390/s21020526 PMID:33450967

Hou, J. (2021). Online teaching quality evaluation model based on support vector machine and decision tree. *Journal of Intelligent & Fuzzy Systems*, *40*(2), 2193–2203. doi:10.3233/JIFS-189218

Hu, G., Mohammadiun, S., Gharahbagh, A. A., Li, J., Hewage, K., & Sadiq, R. (2020). Selection of oil spill response method in Arctic offshore waters: A fuzzy decision tree based framework. *Marine Pollution Bulletin*, *161*, 111705. doi:10.1016/j.marpolbul.2020.111705 PMID:33022490

Li, L., Dai, S., Cao, Z., Hong, J., Jiang, S., & Yang, K. (2020). Using improved gradient-boosted decision tree algorithm based on Kalman filter (GBDT-KF) in time series prediction. *The Journal of Supercomputing*, *76*(9), 6887–6900. doi:10.1007/s11227-019-03130-y

Lou, D., Yang, M., Shi, D., Wang, G., Ullah, W., Chai, Y., & Chen, Y. (2021). K-Means and C4.5 decision tree based prediction of long-term precipitation variability in the Poyang lake basin, China. *Atmosphere (Basel)*, *12*(7), 834. doi:10.3390/atmos12070834

Mao, L., & Zhang, W. (2021). Analysis of entrepreneurship education in colleges and based on improved decision tree algorithm and fuzzy mathematics. *Journal of Intelligent & Fuzzy Systems*, *40*(2), 2095–2107. doi:10.3233/JIFS-189210

Pham, Q. B., Chandra Pal, S., Chakrabortty, R., Saha, A., Janizadeh, S., Ahmadi, K., Khedher, K. M., Anh, D. T., Tiefenbacher, J. P., & Bannari, A. (2021). Predicting landslide susceptibility based on decision tree machine learning models under climate and land use changes. *Geocarto International*, *37*(25), 7881–7907. doi:10.1080/10106049.2021.1986579

Shang, Y., Han, Z., Qiao, Y., & Zhou, J. (2020). Visualization analysis of the journal of intelligent & fuzzy systems (2002–2018). *Journal of Intelligent & Fuzzy Systems*, *38*(3), 2979–2989. doi:10.3233/JIFS-18326

Strain, T., Wijndaele, K., Garcia, L., Cowan, M., Guthold, R., Brage, S., & Bull, F. C. (2020). Levels of domain-specific physical activity at work, in the household, for travel and for leisure among 327 789 adults from 104 countries. *British Journal of Sports Medicine*, *54*(24), 1488–1497. doi:10.1136/bjsports-2020-102601 PMID:33239355

Tao, Y., Guo, S., Shi, C., & Chu, D. (2019). User behavior analysis by cross-domain log data fusion. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 400–406. doi:10.1109/ACCESS.2019.2961769

Xie, W., She, Y., & Guo, Q. (2021). Research on multiple classification based on improved SVM algorithm for balanced binary decision tree. *Scientific Programming*, *2021*, 1–11. doi:10.1155/2021/5560465

Yu, Y. (2020). Research on current situation and development of marine leisure sports in Shandong Peninsula. *Journal of Coastal Research, 115*(SI), 123–126. 10.2112/JCR-SI115-037.1

Yuan, N. (2020). Application of machine learning and cloud computing in social media behavior analysis. *Journal of Intelligent & Fuzzy Systems*, *39*(2), 1831–1842. doi:10.3233/JIFS-179955

Zheng, Y. (2018). Research on the competitiveness of China's leisure sports industry based on statistical method. *Journal of Intelligent & Fuzzy Systems*, *35*(3), 2855–2860. doi:10.3233/JIFS-169639