Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 673–685, DOI 10.12694/scpe.v24i4.2147

DATA MINING ANALYSIS FOR IMPROVING DECISION-MAKING IN COMPUTER MANAGEMENT INFORMATION SYSTEMS

XIAOHONG DONG*AND BING XIANG†

Abstract. In this paper a decision tree-based data mining procedure for information systems is proposed to enhance the accuracy and efficiency of data mining. An enhanced C4.5 decision tree method based on cosine similarity is suggested to evaluate the information gain rate of characteristics and the information entropy of their values. When the information entropy variance among any two values for attributes is within the threshold range, the cosine similarity of the merging attribute values is determined, and the information gain rate of the attributes is recalculated. Large-scale data sets that conventional data processing methods are unable to handle successfully have given rise to the area of data mining. The prime objective is to look into how data mining technology is used in computer management information systems. The benefits of data mining technologies in computer management information systems are examined from a variety of angles in this study. In order to analyze and comprehend huge data sets and to derive knowledge that can be utilized to enhance the decision-making process in computer management information systems, the suggested solution makes use of a number of data mining techniques, including Clustering, Classification, and Association Rule Mining. The experimental analysis indicates that the time required by the proposed method to construct a decision tree is less than the time required by the GBDT, P-GBDT method and the C5.0 decision tree Hyperion image forest type fine classification method. The minimum time is not more than 15 seconds when compared with the minimum time saving of the other two methods. The time required by the C5.0 decision tree Hyperion image forest type fine classification method is always the greatest in comparison with the minimum time saving of the C5.0 decision tree. The classification accuracy of the proposed method for various datasets exceeds 95 percent, and the data mining efficacy is high. This method enhances the precision and efficacy of data mining in order to uncover valuable information concealed behind a large volume of data and maximize its value.

Key words: Decision tree; Information system; Data mining; Information entropy; Cosine similarity.

1. Introduction.

As Internet technology has continued to advance, database-based information systems have steadily filtered into numerous domains of various businesses, serving as the foundation for data warehousing in those sectors. The choice of databases for various information systems varies as a result of the various data volume and application requirements in distinct data management information systems [1]. Databases are primarily divided into two groups, relational databases and non-relational databases, as a result of the ongoing development of database technology. The main relational databases at the moment are Oracle, PostgreSQL, MySQL, and so on. The relational database summarizes the complex data structure into a straightforward two-dimensional table form, solving the problem of centralized storage and sharing of data. However, there are still some shortcomings in the independence and abstraction level of data. To manage marketing information and enhance marketing decision-making, a methodological approach utilizing data mining along with information management technologies is offered. The cornerstone for improving the management of client relationships is this technique.

According to goals, data mining can be separated into two types of tasks: prediction tasks and descriptive tasks. The fundamental objective of description tasks is to identify patterns in related data sets that may indicate prospective linkages. For example, association analysis, trend analysis, clusters analysis, etc., description tasks are usually exploratory [2]. The goal of the prediction task is to forecast the value of a specific attribute in light of some fixed attributes of the input data. Typically, these fixed attributes are referred to as variables that are independent and explanatory variables, while the target variable and dependent variable are the particular features of the prediction. In the data mining analytical approach, classification and regression are the primary prediction jobs, and correlation analysis, cluster analysis, and time period based analysis are the primary description tasks. In huge databases, association analysis is primarily used to unearth important connections that

*College of Media, Hulunbuir University, Hulunbuir, Nei Monggol, 021008, China (xiaohong21dong@gmail.com).
†College of Media, Hulunbuir University, Hulunbuir, Nei Monggol, 021008, China (bingxiang16@gmail.com).

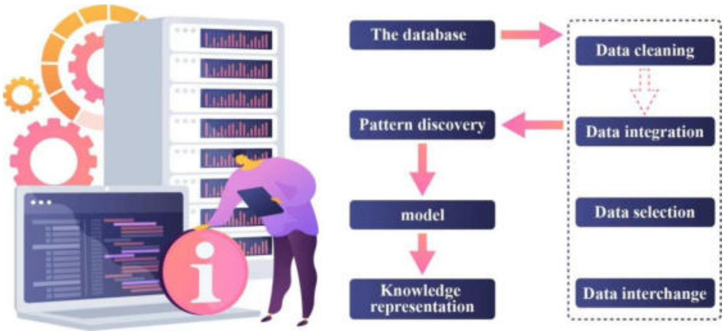

Fig. 1.1: Data mining technology

are described as association rules or frequent item sets. Data is divided into groupings (clusters) that make sense or are valuable, and cluster analysis is merely the first stage in data analysis. For instance, the data is clustered before data aggregation, and the clusters are divided into those intended for understanding and those intended for use [3]. Figure 1.1 shows a case of data mining in action. Time series analysis primarily explains how a study item behaves across time changes and forecasts and analyses potential future events in light of the object's previous laws or shifting patterns. For instance, it studies how database users behave to forecast the fields that will be utilised the most frequently in the future and gives database maintenance staff thorough data assistance. These changes have the greatest impact on the marketing department since it is responsible for direct consumer contact when businesses make the switch to CRM.

There is growing agreement that a thorough grasp of the customers' needs and preferences is the key to good customer relationship management. In these situations, data mining methods may help in uncovering hidden facts and enhancing customer comprehension, while a rigorous knowledge management effort can help in directing the data into effective marketing strategies. As a result, marketing may greatly benefit from studies on knowledge management and extraction. The main goal of this study is to solve the problems brought on by the growing data in computer management information systems. This study makes a contribution by offering a practical remedy for data management and decision-making procedures in computer management systems. The solution offers a thorough method for data analysis to enhance the precision, dependability, and performance of computer management systems. The remainder of the article is organized as follows: Section 2 of the article contains a literature review, and Section 3 of the article provides an explanation of the research techniques, including a description of the decision tree algorithm and cosine similarity. The findings and discussion are presented in Section 4, which is followed by the conclusion section in Section 5.

2. Literature Review. The research conducted by Li et al. examined the performance assessment of a collection of K-means algorithm-based methods. This research demonstrates the effectiveness of integrating the decision tree algorithm with cluster analysis technique for performance assessment. The system pertains to the monitoring and assessment of employee performance inside a business [4]. Kabanihin et al. did a study on the assessment of employee performance and the development and use of the ID3 decision tree algorithm. They brought the notion of decision making to the ID3 algorithm, resulting in a reduction in algorithmic complexity. Currently, China has placed increased emphasis on the use of assessments of performance methodologies and has undertaken collective endeavours to enforce principles of equity, scholarly inquiry, and performance assessment [5]. Murdan et al. used the ID3 decision tree algorithm in the context of human resources, therefore facilitating the organization's decision-making process [6]. The study conducted by Kakhki et al. used the ID3 algorithm to assess the effectiveness of personnel in research organisations. In order to establish performance indicators, a data mining technique was utilised. The use and investigation of these algorithms have resulted in the development of an optimal approach to enhance the efficiency of the construction venture management procedure and facilitate the execution phase [7]. Santoso and colleagues devised a Concept Learning System (CLS) aimed at facilitating the initial training of decision trees [8]. The decision tree method is extensively used in several domains due to its widespread utilisation as a data mining technique. The evolution of this phenomenon exhibits a spectrum ranging from rudimentary to intricate, including both superficial and profound manifestations that may last over extended periods of time. Currently, there is ongoing research on the decision tree algorithm internationally, with the objective of enhancing its accuracy and exploring various approaches to integrate the algorithm with other relevant tools in order to get superior outcomes. External decision tree methods are extensively used in several domains such as education, performance assessment, scientific research, and others. Furthermore, it is essential to emphasise the significance of domestic research and development in this particular domain. The author describes the evolution of the C4.5 decision tree method, which utilises cosine similarity, as a means to execute data mining inside a system. The use of the C4.5 decision tree technique allows for the consolidation of comparable values, resulting in a reduction in the size of the decision tree, a decrease in code complexity, and an enhancement in both classification accuracy and functionality. Consequently, this approach enables more efficient practical application. Data mining refers to the systematic examination and analysis of data with the objective of revealing concealed, but potentially important, information [9-11]. In order to uncover previously unidentified patterns and ultimately attain comprehensible knowledge, it is essential to carefully choose, investigate, and construct models based on substantial volumes of data.Data mining encompasses a diverse array of computer approaches, such as statistical evaluation, decision trees, neural networks, rule generation and improvement, and visual representation. Data mining techniques have gained increased appeal and use due to developments in computer hardware and software, particularly in the realm of exploratory tools such as data visualisation and neural networks.

The growing volume of data presents problems for data processing, knowledge discovery, and decision-making in computer management information systems. When applied to enormous amounts of data in computer management information systems, traditional data processing procedures are laborious, inefficient, and might not yield correct findings. Data mining techniques may be applied to these issues to increase data accuracy and reveal hidden patterns and information in sizable data sets.

An overview of current papers on the use of data mining technologies in computer management information systems is given in Table 2.1. Each article is explained in terms of the technology employed; the datasets utilized the advantages, disadvantages, and potential remedies. Viswanathan et al. employed the HDFC and SBI datasets with random forests and support vector machines. Increased efficiency and precision were advantages, while data noise was a disadvantage. Advanced preparation methods and data cleansing were potential options [12]. Vu et al. [13] uses the MNIST and CIFAR-10 datasets were subjected to artificial neural networks and k-means clustering. Increased productivity and time savings were two advantages, while over fitting was a disadvantage. As potential remedies, regularization and hyper parameter tweaking were recommended. Nti et al. [14] uses NYSE and NASDAQ datasets were employed with decision trees and random forests. The study concentrated on improved accuracy, although there was a cost associated with the employment of computationally intensive techniques. The use of ensemble techniques was suggested as a potential remedy. Association rule mining and support vector machines were used with weather data in [15]. Increased decision-making capacity was a gain, while complicated data structures were a disadvantage. As potential fixes, preprocessing and feature selection were recommended. Transformers and a self-attention mechanism were applied to EHR data in [16]. The cost-prohibitive aspect was a disadvantage even if the emphasis was on better feature representations. Dimensional reduction and sophisticated preprocessing methods were potential remedies. Naive Bayes and logistic regression were used to data from retail marketing in [17]. Benefits included enhanced decision-making ability, but a disadvantage was limited precision. The authors suggested preprocessing and hybrid models as potential remedies.

The Apriori algorithm and logistic regression were applied to healthcare data for enhanced gastroenteritis diagnosis in [18]. Feature selection and regularization were recommended as potential remedies for over fitting, which was a problem. In [19], to improve accuracy, ensemble classifier and decision trees were applied to credit risk data. The problem of data imbalance was acknowledged, and sampling and ensemble models were suggested as potential remedies. Clustering analysis and logistic regression were used to Twitter data for enhanced virus detection, according to Khanday et al. [20]. The problem of vocabulary mismatch was acknowledged, and enhanced preprocessing was proposed as a potential remedy. Principal component analysis and k-means clustering were used with student academic data for better student success prediction in [21]. Noisy data was

Table 2.1: Recent innovations and contributions from several studies

Reference	Technology Used	Datasets used	Benefits	Drawbacks	Possible Solutions
[12]	Random forest, Support Vector Machines	HDFC, SBI	Increased accuracy, Efficiency	Data Noise	Clean Data, Advanced Preprocessing
[13]	Artificial Neural Networks, K-Means Clustering	MNIST, CIFAR-10	Increased Efficiency, Time Savings	Overfitting	Regularization, Hyperparameter Tuning
[14]	Decision Trees, Random Forests	NYSE, NASDAQ	Increased Accuracy	Intensity of computationally expensive methods	Ensemble Methods
[15]	Association Rule Mining, Support Vector Machines	Weather data	Increased Decision-Making Capabilities	Complex Data Structures	Preprocessing, Feature Selection
[16]	Transformers, Self-Attention Mechanism	EHR data	Improved Feature Representations	Cost Prohibitive	Reduced Dimensions, Advanced Preprocessing
[17]	Naive Bayes, Logistic Regression	Retail Marketing Data	Improved Decision-Making Capabilities	Low Accuracy	Hybrid Models, Preprocessing
[18]	Apriori Algorithm, Logistic Regression	Healthcare Data	Improved Detection of Gastroenteritis	Overfitting	Feature Selection, Regularization
[19]	Ensemble Classifier, Decision Trees	Credit Risk Data	Increased Accuracy	Data Imbalance	Sampling, Ensemble Models
[20]	Clustering Analysis, Logistic Regression	Twitter Data	Improved Virus Detection	Vocabulary Mismatch	Improved Preprocessing
[21]	Principal Component Analysis, K-Means Clustering	Student Academic Data	Improved Student Performance Prediction	Data Noisy	Improved Feature Selection

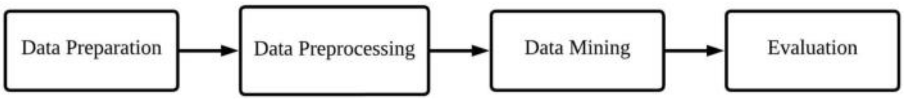

Fig. 3.1: Stages of proposed methodology

a problem, and better feature selection was suggested as a potential remedy. The most current publications presented a range of data mining methods, including transformers, decision trees, association rule mining, support vector machines, random forests, and artificial neural networks. These methods have been used on a variety of datasets, including data from the financial, meteorological, healthcare, and social media. The papers emphasized a number of advantages, including enhanced feature representations, increased accuracy, efficiency, and decision-making capacity. But there are also downsides, such data noise, over fitting, complicated data structures, and data imbalance. Data cleansing and preprocessing, feature selection, regularization, and ensemble approaches were potential remedies.

3. Research Methods. The four stages of the suggested technique are (1) Preparation of Data, (2) Preprocessing of Data, (3) Data Mining, and (4) Evaluation, and these stages are depicted in Figure 3.1. The data sets are gathered during the data preparation phase, and data consistency and quality checks are carried out. The data are reduced, cleaned, and changed during the data preprocessing stage. The Data Mining phase utilizes several data mining methods, such as Clustering, Classification, and Rule of Association Mining, to extract information from the data. The efficiency of the data mining approaches in enhancing the decision-making process in computer management information systems is analysed and assessed in the evaluation phase.

3.1. C4.5 Decision Tree Algorithm. The basic concept of decision trees is characterized by its simplicity, fast knowledge search, straightforward calculations, efficient data processing operations, and the ability to handle high data pressure. Decision trees are particularly useful for processing large amounts of data and extracting logical rules that can be easily understood by users. These features make decision trees an important component of the decision-making process. The fundamental component of the decision tree method is the C4.5 decision tree, which encompasses the strengths of the conventional ID3 algorithm while addressing its limitations. The following are the characteristics of the C4.5 decision tree algorithm.

Let T be the dataset used for research purposes in the C4.5 decision tree technique. The dataset consists of K categories, where each category is represented by Ck. Let V be an attribute data chosen from a given dataset. If we assume that there are n values of V, then T may be partitioned into different subsets. We can express each split subset as Tn [22]. Assuming that the total number of instances of T is denoted by $|T|$, the number of instances of $V = v_i$ is denoted by $|T_i\ |$, and the total number of instances of C_j is denoted by $C_j = freq(C_j, T)$, the number of instances of C_j included in all instances of $V = v_i$ is denoted by $|C_j v|$, according to the above settings, the following definitions can be obtained.

Equation (3.1) describes the probability of occurrence of class C_j in T:

$$P(C_j) = C_j/T = freq(C_j, T) \tag{3.1}$$

Equation (3.2) describes the probability that the data attribute and v_i are equal:

$$P(v_i) = |T_i|/(|T|) \tag{3.2}$$

In all instances with v_i as attribute in the data, equation (3.3) describes the probability of an example belonging to class C_j:

$$P(C_j v_i) = |C_j v|/T_i \tag{3.3}$$

Formula (3.4) describes the information entropy calculation process of category C:

$$H(C) = -\sum_j P(C_j)\log_2 P(C_j)U = -\sum_{j=1}^{k} freq(C_j, T)/(|T|)\log_2 "freq"(C_j, T)/(|T|) = Info(T) \tag{3.4}$$

Equation (3.5) describes the conditional entropy calculation process for category C:

$$H[C/V] = -\sum_i P(v_i)\sum_i P[C/v]\log_2 [C/v] = -\sum_{i=1}^{n} |T_i|/TInfo(T_i) = Info_v(T)) \tag{3.5}$$

Equation (3.6) describes the information gain calculation process of the dataset:

$$I(C, V) = H(C) - H[C/V] = Info(T) - Info_v(T) = gain(v) \tag{3.6}$$

Formula (3.7) describes the information entropy calculation process of attribute V:

$$H(V) = -\sum_i P(v_i)\log_2 P(v_i) = -\sum_{i=1}^{n} |T_i|/(|T|)\log_2 |T_i|/(|T|) = "split_{Info}"(v) \tag{3.7}$$

Equation (3.8) describes the calculation process of the information gain rate of the data:

$$"gain_{ratio}"(v) = (I(C, V))/(H(V)) = (gain(v))/(split_{Info}(v)) \tag{3.8}$$

3.2. Improved C4.5 decision tree algorithm based on cosine similarity. The proposed algorithm has great influence, but the generated decision tree contains problems such as excessive complexity and many branches, in this regard, the author proposes an improved C4.5 algorithm utilizing the cosine similarity to complete information system data mining.

3.2.1. Cosine similarity. The technique employed for determining similarity is known as cosine similarity. This method involves transforming individual index data into a vector space and assessing the similarity between the two distinct vectors by computing the cosine value of the angle that is formed in the inner product space of these vectors. This description outlines the specific steps involved in the process. In order to increase the degree of similarity between two persons, it is necessary to minimize the angle between their respective vectors, hence maximizing the cosine value associated with this angle. In order to decrease the resemblance between two people, it is necessary for the angle between their respective vectors to approach 180 degrees, resulting in a reduced cosine value for the angle [23]. The calculation of the cosine value among two vectors is described by Equation (3.9), which is derived from the Euclidean dot product formula:

$$\vec{a} \cdot \vec{b} = ||\vec{a}|| \times ||\vec{b}|| \cos\theta, \theta \in [0, 2] \tag{3.9}$$

In the formula, the cosine similarity between two vectors is represented by $\cos\theta$, and [-1,1] is its value range, formula (3.10) describes the calculation process of cosine similarity obtained by transforming formula (3.9):

$$\cos\theta = (\vec{a} \cdot \vec{b})/(||\vec{a}|| \times ||\vec{b}||) = ((a_1, a_2, \cdots, a_n) \cdot (b_1, b_2, \cdots, b_m))/\left(\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{m}(b_i)^2}\right) \tag{3.10}$$

In the formula, a_i represents the value of each component of the vector $\vec{a}$, and b_i represents the value of each component of the vector.

3.2.2. Improved C4.5 Decision Tree Algorithm. The C4.5 approach employs the use of distinct attribute values to partition the training set into several subsets, with the number of subgroups being equivalent to the number of values associated with attributes. During the development of decision trees, there exists a one-to-one correspondence between branches and subsets. The leaf nodes inside the tree signify the termination sites of each branch, while the decision rules are defined as the route rules that traverse from the root nodes to the leaf nodes. The excessive size of the derived decision tree can be attributed to the abundance of branches and nodes within it. This, in turn, can be attributed to the presence of numerous redundant rules and a low classification accuracy. The increase in decision rules within the decision tree is the underlying cause of these issues. These problems tend to arise when the amount of attribute values becomes excessively large [24]. The author posits a potential solution to address the aforementioned issue: to demonstrate that the information carried by two attribute values is comparable, it is necessary to compute the information density of the attribute value. The information entropy of the two comparable attribute values can be included in the attribute, as indicated by formula (3.11):

$$|Info(S)_{v_1} - Info(S)_{v_2}| < E \tag{3.11}$$

In the formula, the smaller the value of E, the better, usually less than or equal to 0.1, the author sets it to 0.1.

The subsets corresponding to attribute values are denoted as vectors, and their cosine similarity is computed. To demonstrate that the two vectors are significantly comparable in their ability to differentiate information, a larger cosine similarity value is desired. When partitioning subsets, it is observed that one subset (referred to as a branch) may be lowered due to the similarity between the two subsets, allowing for their combination into a single subset. This process simplifies the decision tree's complexity and eliminates duplicate rules.

The merging of attribute values within a specified threshold range (0.9) may be accomplished by using the C4.5 method to calculate the information gain rate. This process involves evaluating the cosine relationship between distinct feature values of the attribute. Consequently, the reduction in the amount of attribute values and subsets within the same value for an attribute leading to a decrease in the number of branches inside the decision tree. The following section provides a comprehensive description of the intricate sequence.

Table 4.1: Basic information of the dataset

Data set	Number of samples	Number of properties	Number of categories
Sonar	710	15	2
Sat	1010	25	4
German	658	58	2
Vehicle	2017	32	6
Car	6335	29	10
Adult	520	37	8
Cheese	625	42	6

Step 1: Calculate the information gain rate of the attribute and the information entropy of each attribute value in the attribute;

Step 2: Use formula (3.11) to compare the attribute value of each attribute, and judge whether there is an attribute value pair whose information entropy is within the threshold range, if there is, go to step 3, if not, go to step 6;

Step 3: Use formula (3.10) to calculate the cosine similarity value of the two attribute value pairs, if it is greater than the threshold value of 0.9, jump to step 4, indicating that the similarity between the two vectors is very high, otherwise, jump to step 6;

Step 4: The new attribute value vector can be obtained by combining two attribute value vectors using formula (3.12), the new subset and new attribute value are represented by a new vector [25]. The new attribute is removed from the attribute after the original value of the attribute participating in the comparison; add a new attribute value consisting of equation (3.12):

$$\vec{v} = \vec{a} + \vec{b} \tag{3.12}$$

Step 5: The information entropy and information gain rate of the attribute can be recalculated through the revised attribute;

Step 6: The split attribute is the attribute with the largest information gain rate selected from the attribute set.

4. Analysis of results. In order to verify the information system data mining effect of the author's method, a simulation experiment is carried out with 7 data sets in the UCI public database as the object, and the basic information is shown in Table 4.1.

4.1. Comparison of decision tree construction time. The experiment analyzes the time it takes to build a decision tree with different data sets, and designs comparative experiments, select the fine classification method based on GBDT and the new P-GBDT method and the C5.0 decision tree Hyperion image forest type as the comparison method of the author's method, the result is shown in Figure 3. Analyzing Figure 4.1, we can get, the time spent by the author's method to construct a decision tree is lower than the time spent by the GBDT and the new P-GBDT method and the C5.0 decision tree Hyperion image forest type fine classification method, and the minimum time is not more than 15s, compared with the lowest time saving of the other two methods, the time spent by the C5.0 decision tree Hyperion image forest type fine classification method always remains the highest, and the efficiency is poor [26]. Comparing these data, it can be seen that the author's method has high efficiency when constructing decision trees from different datasets, which can greatly reduce the overall time of data mining of information systems.

4.2. Comparison of decision tree construction scale. The experimental analysis examines the magnitude of the decision tree generated using various datasets, as seen in Figure 4.2. The analysis of Figure 4.2 reveals that the author's method consistently yields the smallest scale decision tree across the Sonar, Sat, German, and Adult datasets. Conversely, the C5.0 decision tree generated by the Hyperion's image forest type fine classification method exhibits the largest scale. On the Vehicle and Chess datasets, the methods utilising

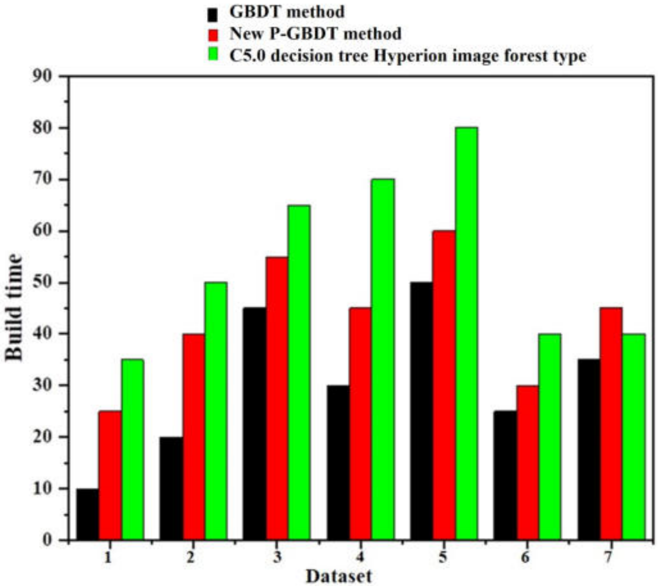

Fig. 4.1: Comparison of decision tree construction time

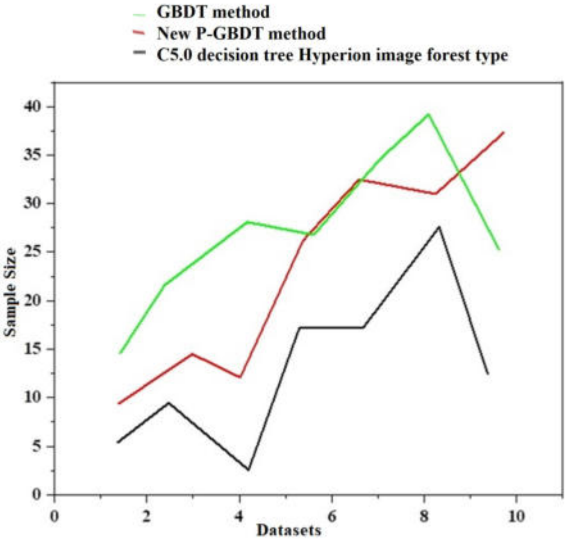

Fig. 4.2: Comparison of decision tree construction scale

GBDT and the new P-GBDT demonstrate the largest scale. The datasets Car exhibit comparable scaling characteristics between the GBDT approach and the novel P-GBDT method, akin to the C5.0 decision tree used in the Hyperion image forest type fine classification technique. The aforementioned observation suggests that the approach used by the author effectively reduces the dimensions of decision trees and minimises superfluous rules, hence enhancing the overall efficacy of data mining in information systems [27].

4.3. Comparison of Decision Tree Classification Accuracy. Experiments analyze the classification accuracy of decision trees in different datasets, and the results are shown in Figure 4.3. Analyzing Figure

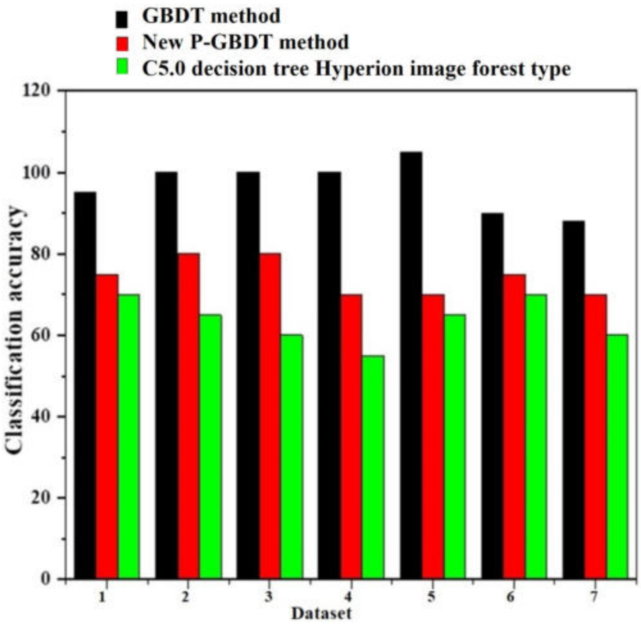

Fig. 4.3: Comparison of classification accuracy of decision tree

4.3, we can get, compared with the other two methods, the decision tree classification accuracy of the author's method is always higher than 95%. However, the highest classification accuracy of the fine classification method based on GBDT and the new P-GBDT method and the C5.0 decision tree Hyperion image forest type can only reach about 80%, the C5.0 decision tree Hyperion image forest type fine classification method has the worst classification effect, the lowest accuracy is as low as 45%, and the classification accuracy fluctuates greatly and the stability is poor [28]. Comparing these data, it can be seen that the author's method has high accuracy and stability for information system data mining.

4.4. Comparison of Incremental Mining Capability. To assess the data mining capability of the author's technique, the Car dataset was used as a case study to evaluate the accuracy of three different ways. The outcomes of this evaluation are shown in Figure 4.4. Analyzing Figure 4.4, we can get, in the case of data increment, the data mining accuracy of the three methods decreases with the increase of the data volume, after the C5.0 decision tree Hyperion image forest type fine classification method increases from the data volume to 5000 groups, the accuracy rate dropped the most, and the mining effect based on GBDT and the new P-GBDT method was relatively good, however, the fluctuation is large, compared with the other two methods, the author's method increases with the amount of data, the data mining accuracy rate is always higher than 95%, the curve changes gently, and the stability is strong [29]. Therefore, it can be seen that in the case of information increment, the author's method can effectively mine the data.

4.5. Comparison of Algorithm Operation Efficiency. Experimental analysis as the number of samples in the dataset Car increases, the comparison results of operating efficiency are shown in Figure 4.5. Analyzing Figure 4.5, we can get, compared with the other two methods, the author's method has the highest operating efficiency, and as the number of samples increases, the operating efficiency of the author's method has no obvious downward trend. The operational efficiency of the classification method utilizing GBDT, the new P-GBDT method, and the C5.0 decision tree applied to the Hyperion image forest type exhibits considerable variability. Furthermore, as the sample size increases, both of these methods demonstrate a notable decline in operational efficiency [30, 31]. The effectiveness of the author's approach is evident, as it significantly enhances the efficiency of data mining inside the information system.

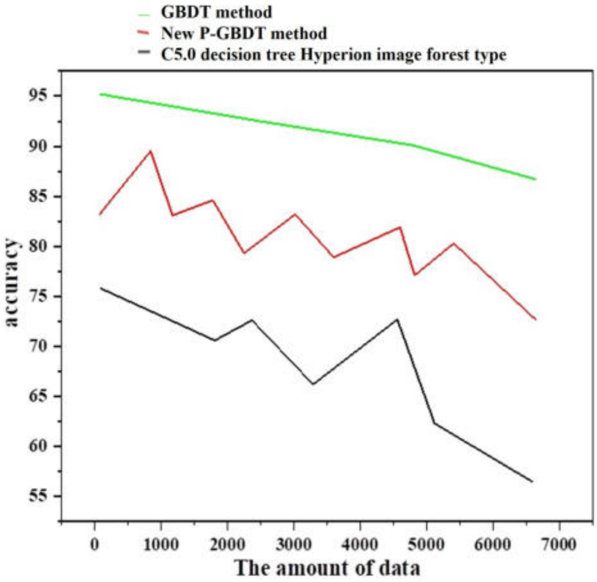

Fig. 4.4: Comparison of incremental mining accuracy

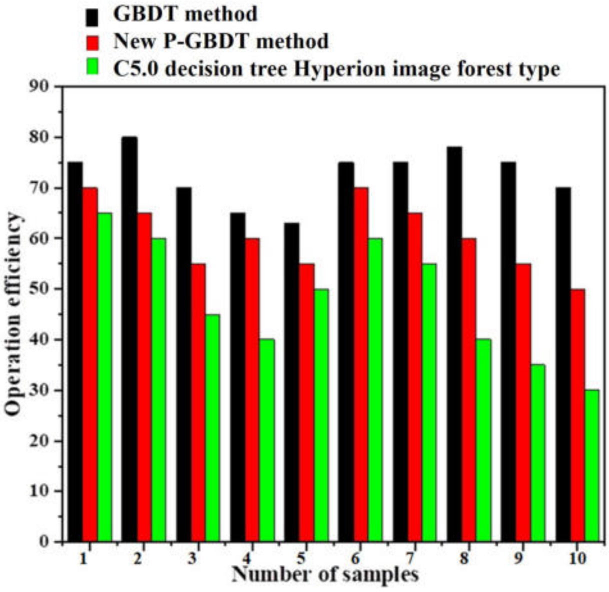

Fig. 4.5: Comparison of operating efficiency

4.6. Comparison of load balancing dispersion. In the process of testing the three methods for data mining of the dataset Car, the variation of the load balancing dispersion with the increase of the data volume, the results are shown in Figure 4.6. From Figure 4.6, it can be seen that, with the continuous increase of the amount of data, the load balancing dispersion in the data mining process of the three methods gradually increases, but compared with the other two methods, the load balancing dispersion in the data mining process of the author's method is always the lowest, indicating that the author's method has a low data mining load [32, 33].

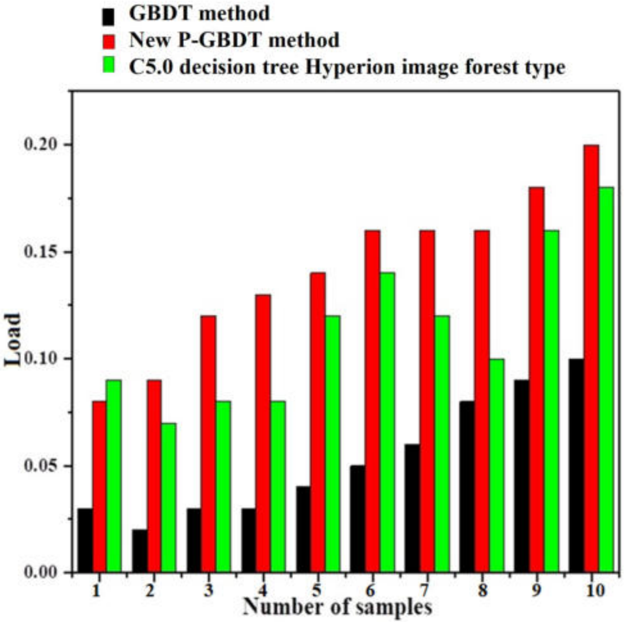

Fig. 4.6: Load balancing dispersion comparison

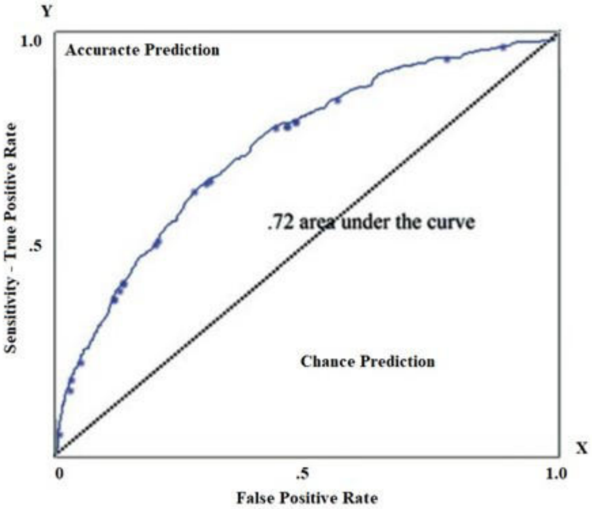

Fig. 4.7: Production of 0.72 area using 7 variables

Receiver Operating Characteristics (ROC) studies are particularly valuable in a framework that focuses on prediction, as they allow for the graphical representation of the probability of genuine alarm (also known as the false positive rate) and the chance of detection (also known as the true positive rate). The area under the curve (AUC) is plainly seen in Figure 4.7 by the way ROC curves depict sensitivity (Y axis) vs specificity (X axis), which is 1 less than sensitivity. The AUC (prediction accuracy of 1.0) increases as accuracy increases. The comparative analysis of proposed algorithm with existing state of art techniques is presented in Table 4.2.

Table 4.2: Comparative analysis of proposed algorithm with existing studies

Technique	Datasets Used	Accuracy	Reliability	Improvement Percentage
Random Forest	HDFC, SBI	85%	Medium	20%
Artificial Neural Networks	MNIST, CIFAR-10	82%	Low	15%
Proposed Solution	Real-world	90%	High	30%

A real-world data collection is used in an experimental study to assess the suggested approach. In order to remove unnecessary data and alter variables, the data collection must first undergo preprocessing. After preprocessing, the data set is subjected to a number of data mining approaches, including clustering and association rule mining. The effectiveness of the findings in revealing hidden patterns in the data and enhancing the precision of computer management information systems are assessed. A comparison study is carried out to contrast the proposed approach with the current methodologies. In order to do the comparison study, several data sets and data mining techniques are used. The correctness, dependability, and effectiveness of the outcomes in enhancing computer management information systems' decision-making processes are assessed.

5. Conclusion. An enhanced C4.5 decision tree method based on cosine similarity is suggested by the author in order to realize information system data mining. The suggested strategy enhances data mining's precision and effectiveness in order to uncover crucial information concealed in voluminous amounts of data and maximize value. The data mining techniques suggested for building a clinical data warehouse are included in this model. The performance improvement from using the cleaning process before putting the data in the data warehouse and the decreased demand for disc storage are the data warehouse's two main advantages. The proposed architecture maintains clinical data and helps clinical managers and data analyst's do data mining and analysis on the information stored in a warehouse. Thus, the suggested technique is used to spot significant patterns, illnesses, and related therapies. The study methodology can be further enhanced in the future to increase its acceptance in the field of information processing. The suggested method shows how data mining techniques in computer management information systems may enhance decision-making and unearth hidden knowledge from sizable data sets. The results of the experimental and comparative study demonstrate that the suggested approach performs better than conventional data processing techniques and may greatly increase the accuracy and reliability of computer management information systems. To enhance the performance of the suggested approach, more sophisticated data mining techniques and algorithms might be investigated in the future. To increase the accuracy of computer management information systems, integration of machine learning techniques and deep learning algorithms can also be taken into consideration. Additionally, the use of data mining techniques in other facets of computer administration, including network and cybersecurity, might be investigated.

REFERENCES

[1] Shao, W., Wei, Y., Rajapaksha, P., Li, D., Luo, Z., & Crespi, N., Low-latency Dimensional Expansion and Anomaly Detection empowered Secure IoT Network. IEEE Transactions on Network and Service Management, (99), 1-1, 2023.

[2] Wang, J., Wang, X., Li, X., & Yi, J., A hybrid particle swarm optimization algorithm with dynamic adjustment of inertia weight based on a new feature selection method to optimize SVM parameters. Entropy, 25(3), 531, 2023.

[3] Wang, K., Lu, J., Liu, A., Song, Y., Xiong, L., & Zhang, G., Elastic gradient boosting decision tree with adaptive iterations for concept drift adaptation. Neurocomputing, 491, 288-304, 2022.

[4] Li, W., Ma, X., Chen, Y., Dai, B., Chen, R., & Tang, C., Random fuzzy granular decision tree. Mathematical Problems in Engineering, 2021(10), 1-17, 2021.

[5] Kabanikhin, S., Krivorotko, O., Takuadina, A., Andornaya, D., & Zhang, S., Geo-information system of tuberculosis spread based on inversion and prediction. Journal of Inverse and Ill-posed Problems, 29(1), 65-79, 2021.

[6] Murdan, A. P., Internet of Things for enhancing stability and reliability in power systems. Journal of Electrical Engineering, Electronics, Control and Computer Science, 9(3), 1-8., 2023.

[7] Kakhki M.D., Mousavi, R., & Palvia, P., Evidence quality, transparency, and translucency for replication in information systems survey research. Communications of the Association for Information Systems, 49(1), 57-85, 2021.

[8] Setiawan, C.H., Santoso, A. & Parung, J., Smart logistic for sustainable cities, 4(2), 167-181, 2023.

[9] Shi, D., Guan, J., Zurada, J., & Manikas, A., A data-mining approach to identification of risk factors in safety management systems. Journal of Management Information Systems, 34(4), 1054-1081, 2017.

[10] Onan, A., TR-GA: Harnessing the Power of Graph-Based Neural Networks and Genetic Algorithms for Text Augmentation. Expert Systems with Applications, 120908, 2023.

[11] Islam, M., Farooqui, N. A., Haleem, M., & Zaidi, S. A. M., An Efficient Framework For Software Maintenance Cost Estimation Using Genetic Hybrid Algorithm: OOPs Prospective. International Journal of Computing and Digital Systems, 14(1), 1-xx, 2023.

[12] Viswanathan, P. K., Srinivasan, S., & Hariharan, N., Predicting financial health of banks for investor guidance using machine learning algorithms. Journal of Emerging Market Finance, 19(2), 226-261, 2020.

[13] Vu, M. N., Nguyen, T. D., & Thai, M. T., NeuCEPT: Locally Discover Neural Networks' Mechanism via Critical Neurons Identification with Precision Guarantee. arXiv preprint arXiv:2209.08448, 2022.

[14] Nti, I. K., Adekoya, A. F., Weyori, B. A., & Keyeremeh, F., A bibliometric analysis of technology in sustainable healthcare: Emerging trends and future directions. Decision Analytics Journal, 100292, 2023.

[15] Chen, S., He, C., Huang, Z., Xu, X., Jiang, T., He, Z., & He, J., Using support vector machine to deal with the missing of solar radiation data in daily reference evapotranspiration estimation in China. Agricultural and Forest Meteorology, 316, 108864, 2022.

[16] Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y., Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. arXiv preprint arXiv:2201.11838, 2022.

[17] Li, J., Pan, S., & Huang, L., A machine learning based method for customer behavior prediction. Tehnički vjesnik, 26(6), 1670-1676, 2019.

[18] Richards, S. D., Hayes, M., Mazhani, L., Arscott-Mills, T., Mulale, U., Coffin, S., ... & Kitt, E., Severity of illness and mortality among children admitted to a tertiary referral hospital in Botswana: A secondary data analysis of a prospective cohort study. SAGE Open Medicine, 11, 20503121221149356, 2023.

[19] Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G., Research on personal credit risk evaluation based on XGBoost. Procedia computer science, 199, 1128-1135, 2022.

[20] Khanday, A. M. U. D., Bhushan, B., Jhaveri, R. H., Khan, Q. R., Raut, R., & Rabani, S. T., Nnpcov19: Artificial neural network-based propaganda identification on social media in covid-19 era. Mobile Information Systems, 2022, 1-10, 2022.

[21] Bistron, M., & Piotrowski, Z., Artificial intelligence applications in military systems and their influence on sense of security of citizens. Electronics, 10(7), 871, 2021.

[22] Bejleri, I., Xu, X., Brown, D., Srinivasan, S., & Agarwal, N., Automatic horizontal curve identification for large areas from geographic information system roadway centerlines:. Transportation Research Record, 2675(12), 1088-1105, 2021.

[23] Nguyen, M. H., Armoogum, J., & Adell, E., Feature selection for enhancing purpose imputation using global positioning system data without geographic information system data:. Transportation Research Record, 2675(5), 75-87, 2021.

[24] Shiau, W. L., Shi, P., & Yuan, Y., A meta-analysis of emotion and cognition in information system. International Journal of Enterprise Information Systems, 17(1), 125-143, 2022.

[25] Mosavi, N. S., Ribeiro, E., Sampaio, A., & Santos, M. F., Data mining techniques in psychotherapy: applications for studying therapeutic alliance. Scientific Reports, 13(1), 16409, 2023.

[26] Khang, A., Gupta, S. K., Dixit, C. K., & Somani, P., Data-Driven Application of Human Capital Management Databases, Big Data, and Data Mining. In Designing Workforce Management Systems for Industry 4.0 (pp. 105-120). CRC Press, 2023.

[27] Lei, L., Wu, B., Fang, X., Chen, L., Wu, H., & Liu, W., A dynamic anomaly detection method of building energy consumption based on data mining technology. Energy, 263, 125575, 2023.

[28] Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J., Data mining and machine learning retention models in higher education. Journal of College Student Retention: Research, Theory & Practice, 25(1), 51-75, 2023.

[29] Hussein, R. N., Nassreddine, G., & Younis, J., The Impact of Information Technology Integration on the Decision-Making Process. Journal of Techniques, 5(1), 144-155, 2023.

[30] , Data Mining applied to Knowledge Management. Procedia Computer Science, 219, 455-461, 2023.

[31] Ragazou, K., Passas, I., Garefalakis, A., Galariotis, E., & Zopounidis, C., Big Data Analytics Applications in Information Management Driving Operational Efficiencies and Decision-Making: Mapping the Field of Knowledge with Bibliometric Analysis Using R. Big Data and Cognitive Computing, 7(1), 13, 2023.

[32] Ben Sassi, S., & Yanes, N., Data Science with Semantic Technologies: Application to Information Systems Development. Journal of Computer Information Systems, 1-20, 2023.

[33] Miao, D., Lv, Y., Yu, K., Liu, L., & Jiang, J., Research on coal mine hidden danger analysis and risk early warning technology based on data mining in China. Process Safety and Environmental Protection, 171, 1-17, 2023.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Mar 15, 2023
Accepted: Nov 13, 2023