DOI: 10.53469/wjimt.2025.08(04).12

On the Application of Data Mining Technology in Today's Era

Zhengde Bao

School of Computer and Software, Jincheng college, Sichuan University, Chengdu, Sichuan, 611731

Abstract: ***In the era of data, almost all the leading figures in the IT industry have their own huge data storage database. The rapid development of informatization is accompanied by the increasing amount of data to be analyzed, the larger the data to be analyzed and the more complex the operation of extracting useful information. This paper analyzes the application of data mining technology.***

keywords: Data Mining; Data Analysis; Data Processing; Data Application.

1. INTRODUCTION

Big data is rapidly awakening in the era of information technology. At the same time, the 'big data era' followed closely behind. As the name suggests, the 'big data era' is a brand new era that comprehensively presents big data. Big data plays a huge role in our daily lives, including food, drink, entertainment, and even in important fields such as healthcare, technology, and military development that are closely related to the country. As a result, the data is expanding and even exploding exponentially. However, the initial large amount of data generated is a complex and chaotic pile of data, and obtaining useful data from it requires certain techniques, in which case data mining plays an important role. Mining useful data and analyzing it to obtain decision-making information will play a crucial role in business, economics, and other fields. The proper use of data mining techniques to obtain the truly needed information can better achieve management goals. Gong et al. [1] developed an ensemble machine learning approach to optimize decision support systems, significantly improving risk assessment capabilities. Cybersecurity has seen notable progress through Bohang et al. [2], who implemented active learning with hyperparameter optimization for enhanced image steganalysis, achieving superior detection accuracy in digital forensics. Industrial applications have benefited from AI integration, as shown by Zhao et al. [3] in their deep learning-based optimization of steel production scheduling, and Yao et al. [4] who innovatively combined drone technology with 3D printing for rapid post-disaster shelter construction. Healthcare applications are particularly promising, with Lin et al. [5] demonstrating how intelligent exercise monitoring can improve executive function in ADHD children, while Peng et al. [6] investigated the relationship between aerobic exercise intensity, cognitive function, and sleep quality. Logistics optimization has advanced through Luo et al. [7]'s novel path planning algorithm integrating Transformer and GCN networks for intelligent robots. Medical AI applications continue to expand, evidenced by Shen et al. [8]'s LSTM-based system for precise anesthetic dosing in cancer surgeries. The transportation sector is being revolutionized by Wang et al. [9]'s end-to-end autonomous driving framework. Security concerns in AI systems are addressed by Liu et al. [10] through their privacy-preserving hybrid ensemble model for network anomaly detection. Finally, Lyu et al. [11] contributed to computer vision by developing optimized CNNs for efficient 3D point cloud object recognition, enabling real-time processing in robotics and autonomous systems.

2. DATA MINING

The era of big data is not only the explosion of massive data, but also promoting the development of various data processing technologies [3]. Compared with earlier data processing techniques, traditional data processing techniques were mostly symbolic of statistical analysis. The key techniques of data mining, which involve clustering, classification, and association analysis, mainly focus on discovering information and knowledge. Obtain the value that data can generate from a large amount of complex and irregular data. Whether it is economic development or commercialization, current data mining technologies are a major optimization of traditional data analysis and processing techniques. With the development of technology, data is becoming increasingly large, and the functions of data processing technology will also become more powerful [4].

3. DATA MINING - CLASSIFICATION ANALYSIS

3.1 Classification analysis

Classification analysis can be simply understood as classifying data, but the categories obtained from classification must have a certain basis, that is, the classified data must have generally similar attributes. Generally speaking, the first step is to find a model that can distinguish and describe data classes, mainly using modeling methods to infer object classes whose class labels are unknown. Data classification processing can also be applied to data forecasting, which can summarize the given data from historical data to illustrate and obtain predictions for unknown data. Classification usually outputs discrete data, which is mostly reflected in the pattern of decision trees. According to the data values you give, it scans from the root of the tree in sequence, and goes up along the branches that meet the data conditions. When it reaches the leaves, it can quickly determine its category. In addition to decision tree analysis methods, support vector machines (SVM) and Bayesian networks are also crucial analysis methods.

3.2 Case Analysis

A certain bank now has a "Customer Credit Card Application Information. xlsx" table, which is used to study the factors related to whether customers can apply for credit cards. The fields in this table are shown in Table 1 below:

Table 1: Customer Credit Card Application Information

Field	Describe	Role
Number	1000~	Not have
Age	10~80	Input
Gender	Male, female	Input
Registered residence	Beijing, shanghai,……	Input
Marital status	Widowed, married, unmarried, divorced	Input
Education level	Junior high school or below, associate's degree, bachelor's degree, master's degree or above, high school	Input
Occupation type	Individual businesses, other enterprises, state-owned enterprises, foreign-funded enterprises, private enterprises	Input
Years of service	0~50	Input
Personal income	10000~9.9E10	Input
Residential type	Self purchased property, rented property, other	Input
Vehicle situation	Have,not have	Input
Insurance payment	Have,not have	Input
Credit situation	Repaying in progress Normal repayment No loan records There are currently no loans available Still in arrears Overdue repayment	Input
Credit rating	A~F	Input
Is the application successful	Success,fail	Target

The target field is' whether the application is successful ', with' number 'as none and the rest as input fields. The Bayesian network model is used as the analysis model. After inputting the data source through the "source" and performing the "type" operation, selecting the "Bayesian network model" can result in the following figure 1:

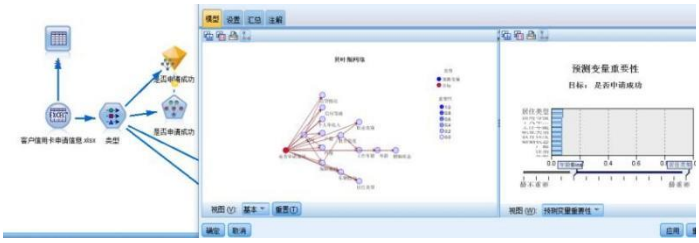


Figure 1: Bayesian network model

The analysis of the Bayesian network model results in Figure 1 shows that the most critical factors affecting whether customers can apply for credit cards are "credit rating" and "type of residence". Therefore, when customers apply for credit cards, banks will focus on their credit rating and type of residence.

4. DATA MINING - CLUSTER ANALYSIS

4.1 Clustering

Cluster analysis is a form of data processing that aggregates information based on its attributes without clear classification analysis. The simple definition of classification and clustering can be that data is classified into different categories based on certain attributes, which is called classification. Clustering is the process of clustering data into different categories based on some characteristics of the data (one by one, with supervised learning for the former and unsupervised learning for the latter). In practical operation, cluster analysis can first perform cluster analysis on a large amount of complex data based on certain characteristics of the data, and then analyze the clustered categories. For example, in a data table of customer call duration, customers can be clustered into business users, large customers, ordinary users, etc. based on their different call types (such as conference calls, home calls, etc.), different call time periods (weekdays, weekends, morning, afternoon, evening, etc.), and different call durations (1-10 hours). Then, different consumer packages can be recommended to users who are clustered into different groups. The commonly used clustering methods include condensed hierarchical clustering, K-means clustering algorithm, etc.

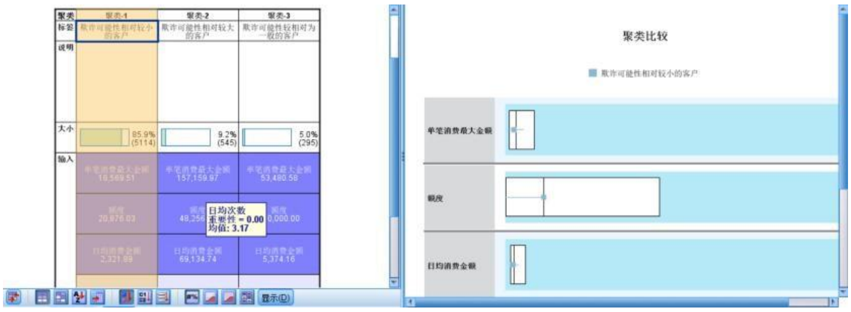
4.2 Case Analysis

A certain bank currently has a table titled 'Is there fraud? Xlsx', which uses cluster analysis to study which customers may engage in fraudulent behavior. Table 2 contains the following fields:

Table 2: Is there fraud present

Field	Describe	Role
Number	1000~	Five
Limit	10000~100000	Input
Daily average consumption amount	30~81797	Input
Daily average frequency	1~28	Input
Maximum amount for a single consumption	30~500000	Input
Personal income - continuous	10416~9.9e10	Input
Is there fraud present	0, 1	Target
Is a single transaction overdrawn	Exceeding, not exceeding	Input
Does daily consumption exceed income	Exceeding, not exceeding	Input
Card swiping frequency	Not frequent, very frequent, frequent	Input

Now, the key fields for operation processing are "limit", "daily consumption amount", "daily frequency", "maximum amount of single consumption", and "personal income". Using the K-means as the analysis model, the data source is inputted through the "source", and then the "type" operation is performed. By selecting the "K-Means model" and setting the number of clusters to "3", the results shown in Figure 2 can be obtained:



聚类	聚类-1	聚类-2	聚类-3
标签	欺诈可能性相对较小的客户	欺诈可能性相对较大的客户	欺诈可能性较相对为一般的客户
说明			
大小	85.9% (5114)	9.2% (545)	5.0% (295)
输入	单笔消费最大金额 16,569.51	单笔消费最大金额 157,159.97	单笔消费最大金额 53,480.58
	额度 20,876.03	额度 48,256	额度 0,000.00
	日均消费金额 2,321.89	日均消费金额 69,134.74	日均消费金额 5,374.16

日均次数
重要性 = 0.00
均值: 3.17

Figure 2: Clustering Model

As shown in the clustering model in Figure 2, based on the indicators of "quota", "daily average consumption amount", "daily average frequency", "maximum amount of single consumption", and "personal income", clusters 1-3 can be divided into customers with high, low, and average fraud likelihood.

5. DATA MINING - ASSOCIATION ANALYSIS

5.1 Association

Association analysis is the process of identifying potential connections among seemingly disorganized data based on certain characteristic attributes inherent in the data itself, and even deriving regularities based on these connections. The association analysis method is used for hidden large data dense and interesting connections. Often, their connections are displayed by association rules or frequent itemsets. The algorithms for correlation analysis include Apriori algorithm and FP growth algorithm.

The central idea of the Apriori algorithm is to generate frequent project teams based on the iterative method of utilizing multiple layers of candidates.

The algorithmic process can be summarized as "merging" and "pruning". Merge: find the last distinct merge, prune: find the subset that is not a frequent itemset and prune it.

The FP growth algorithm is a frequent pattern growth algorithm, which recursively grows frequent patterns and database delineation.

Example: Table 3 below is a transaction set that covers 7 transactions. Each transaction represents a collection of goods purchased by a consumer in a store at once. Using the Apriori algorithm with a minimum support of 30% and a minimum confidence of 80%, find all frequent itemsets and strong association rules?

Table 3: Product List

Transaction Number	Product list
TO01	Beef, chicken, milk
T002	Beef, cheese
T003	Cheese, boots
TO04	Beef, chicken, cheese
T005	Beef, chicken, clothes, cheese, milk
TO06	Chicken, clothes, milk
T007	Chicken, milk, clothes

Answer:

(1) Find all frequent itemsets (with a minimum support of 30%)

1) 1-Item set:

1-Candidate itemset C1 and its supported:

Beef=4/7, chicken=5/7, milk=4/7, Sup {milk}=4/7, Sup {boots}=1/7, Sup {clothes}=3/7

1- Frequent itemset F1:

{Beef}, {Chicken}, {Milk}, {Milk}, {Clothes}

2) 2-Item set:

2-Candidate itemset C2 and its support:

Beef, chicken=3/7, beef, milk=2/7, beef, milk=3/7, beef, clothing=1/7, chicken, milk=4/7, chicken, milk=2/7, chicken, clothing=3/7, milk, milk=1/7, milk, clothing=3/7, Sup {milk, clothes}=1/7

2- Frequent itemset F2:

{Beef, Chicken}, {Beef, Milk}, {Chicken, Milk}, {Chicken, Clothes}, {Milk, Clothes}

3) 3-Item Set:
3-Candidate itemset C3 and its support:

Beef, chicken, milk=2/7, chicken, milk, clothing=3/7

3- Frequent itemset F3:

{Chicken, milk, clothes}

The frequent itemsets obtained from the above are:

1-Item set: {Beef}, {Chicken}, {Milk}, {Milk}, {Clothes}

2-itemsets: {Beef, Chicken}, {Beef, Milk}, {Chicken, Milk}, {Chicken, Clothing}, {Milk, Clothing}

Item Set: {Chicken, Milk, Clothes}

(2) Find all strong association rules (with a minimum confidence level of 80%)

1) The association rule and its confidence level are composed of 2-:

Conf ({Beef} → {Chicken})=3/4, Conf ({chicken} → {beef})=3/5,
Conf ({beef} → {milk})=3/4, Conf ({milk} → {milk})=3/4,
Conf ({chicken} → {milk})=4/5, Conf ({milk} → {chicken})=1,
Conf ({chicken} → {clothing})=3/5, Conf ({clothing} → {chicken})=1,
Conf ({milk} → {clothes})=3/4, conf ({clothes} → {milk})=1

The minimum confidence level of 80% for strong association rules is:

{Milk} → {Chicken}, {Chicken} → {Milk}, {Clothes} → {Chicken}, {Clothes} → {Milk}

2) The association rule and its confidence level are composed of 3-:

Conf ({Milk} → {Chicken, Clothes})=3/4, conf ({Milk, Clothes} → {Chicken})=1, conf ({Chicken} → {Milk, Clothes})=3/5, Conf ({chicken, milk} → {clothing})=3/4, conf ({clothing} → {chicken, milk})=1, conf ({chicken, clothing} → {milk})=1,

The strong association rule with a minimum confidence level of 80% is:

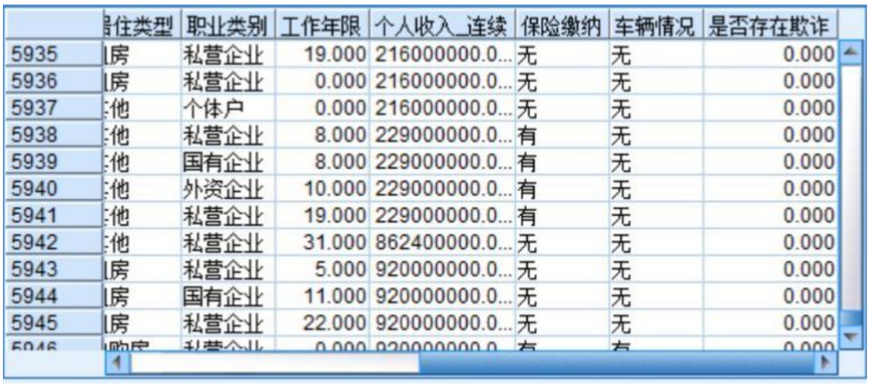
{Clothes} → {Chicken, Milk}, {Chicken, Clothes} → {Milk}, {Milk, Clothes} → {Chicken}

The strong correlation obtained from the above is as follows:

{Milk} → {Chicken}, {Chicken} → {Milk}, {Clothes} → {Chicken}, {Clothes} → {Milk}
{Clothes} → {Chicken, Milk}, {Chicken, Clothes} → {Milk}, {Milk, Clothes} → {Chicken}

5.2 Case Analysis

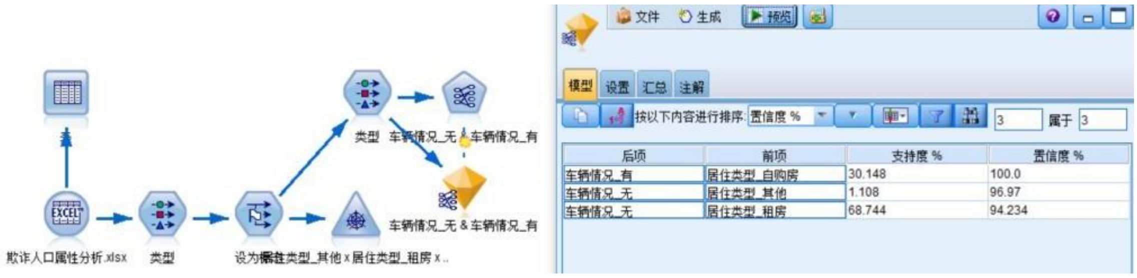
A certain bank currently has a "Fraudulent Population Attribute Analysis. xlsx" table, which includes the fields shown in Figure 3. This table is used to study the correlation between customers' "residential type" and "vehicle situation":



	住类型	职业类别	工作年限	个人收入_连续	保险缴纳	车辆情况	是否存在欺诈
5935	房	私营企业	19.000	216000000.0...	无	无	0.000
5936	房	私营企业	0.000	216000000.0...	无	无	0.000
5937	他	个体户	0.000	216000000.0...	无	无	0.000
5938	他	私营企业	8.000	229000000.0...	有	无	0.000
5939	他	国有企业	8.000	229000000.0...	有	无	0.000
5940	他	外资企业	10.000	229000000.0...	有	无	0.000
5941	他	私营企业	19.000	229000000.0...	有	无	0.000
5942	他	私营企业	31.000	862400000.0...	无	无	0.000
5943	房	私营企业	5.000	920000000.0...	无	无	0.000
5944	房	国有企业	11.000	920000000.0...	无	无	0.000
5945	房	私营企业	22.000	920000000.0...	无	无	0.000
5946	购房	私营企业	0.000	920000000.0...	有	有	0.000

Figure 3: Fraudulent Population Attributes

Now, with "residential type" and "vehicle situation" as the associated fields, the data is input through "source". After operating with "type", "set as flag" is added to select and create the residential type and vehicle situation as flag fields. Then, "type" is added again afterwards to set "vehicle situation _ present/absent" as the target, "residential type _ rental/self purchased/other" as the input. Finally, the "Apriori model" is added to obtain Figure 4:



后项	前项	支持度 %	置信度 %
车辆情况_有	居住类型_自购房	30.148	100.0
车辆情况_无	居住类型_其他	1.108	96.97
车辆情况_无	居住类型_租房	68.744	94.234

Figure 4: Apriori model

Analysis: As shown in Figure 4, the Apriori model, under the conditions of a minimum condition support of 0.0 and a minimum rule confidence (%) of 90.0, can conclude that customers who rent or reside in other types of homes generally do not have vehicles, while customers who own their own homes do. This is also in line with the actual situation.

6. CONCLUSION

The 'data explosion' is the era we live in. The generation of data promotes the development of data analysis and processing technology. Therefore, many data processing and analysis techniques have gradually developed. At this point, it is crucial for leaders of listed companies to select appropriate data analysis and processing techniques, extract deep level information from the vast data, and then refine it to obtain data that truly generates value for the company. Therefore, data mining technology has emerged in the new era and new technologies [5].

REFERENCES

[1] Gong, C., Lin, Y., Cao, J., & Wang, J. (2024, October). Research on Enterprise Risk Decision Support System Optimization based on Ensemble Machine Learning. In Proceeding of the 2024 5th International Conference on Computer Science and Management Technology (pp. 1003-1007).

[2] Bohang, L., Li, N., Yang, J. et al. Image steganalysis using active learning and hyperparameter optimization. Sci Rep 15, 7340 (2025). https://doi.org/10.1038/s41598-025-92082-w

[3] Zhao, H., Chen, Y., Dang, B., & Jian, X. (2024). Research on Steel Production Scheduling Optimization Based on Deep Learning.

[4] Yao, T., Jian, X., He, J., & Meng, Q. (2025). Drone-3D Printing Linkage for Rapid Construction of Sustainable Post-Disaster Temporary Shelters.

[5] Lin, L., Li, N., & Zhao, S. (2025). The effect of intelligent monitoring of physical exercise on executive function in children with ADHD. Alexandria Engineering Journal, 122, 355-363.

[6] Peng, Y., Zhang, G., & Pang, H. (2025). Impact of Short-Duration Aerobic Exercise Intensity on Executive Function and Sleep. arXiv preprint arXiv:2503.09077.

[7] Luo, H., Wei, J., Zhao, S., Liang, A., Xu, Z., & Jiang, R. (2024). Intelligent logistics management robot path planning algorithm integrating transformer and gcn network. IECE Transactions on Internet of Things, 2(4), 95-112.

[8] Shen, Z., Wang, Y., Hu, K., Wang, Z., & Lin, S. (2025). Exploration of Clinical Application of AI System Incorporating LSTM Algorithm for Management of Anesthetic Dose in Cancer Surgery. Journal of Theory and Practice in Clinical Sciences, 2, 17-28.

[9] Wang, Y., Shen, Z., Hu, K., Yang, J., & Li, C. (2025). AI End-to-End Autonomous Driving.

[10] Liu, S., Zhao, Z., He, W., Wang, J., Peng, J., & Ma, H. (2025). Privacy-Preserving Hybrid Ensemble Model for Network Anomaly Detection: Balancing Security and Data Protection. arXiv preprint arXiv:2502.09001.

[11] Lyu, T., Gu, D., Chen, P., Jiang, Y., Zhang, Z., & Pang, H. & Dong, Y.(2024). Optimized CNNs for Rapid 3D Point Cloud Object Recognition. arXiv preprint arXiv:2412.02855.

Author Profile

Zhengde Bao (1989-), male, Han, Harbin, Heilongjiang, graduate student, Jincheng College, Sichuan University, research direction: e-commerce.