The application of data mining technology in enterprise customer business

Haiyue Cheng[1*]
PhD, Marketing (DEA), School of Distance Education, Universiti Sains Malaysia, 11800 Penang,Malaysia
haiyue@student.usm.my
Dr. Siti Haslina Md Harizan[2]
Senior Lecturer, School of Distance Education, Universiti Sains Malaysia, 11800 Penang,Malaysia
sitihaslina@usm.my
Bingyu Fan[3]
PhD, Supply Chain Management, School of Management, Universiti Sains Malaysia, 11800 Penang,Malaysia
fanbingyu@student.usm.my
Shuangshuang Zhang[4]
PhD, School of Management, Universiti Sains Malaysia, 11800 Penang,Malaysia
e27776697shuang@student.usm.my

Abstract: In order to retain current clients and draw in new ones, customer relationship management is crucial. Businesses are becoming increasingly focused on developing relationships with their customers as a result of the implementation of consumer distribution modes. A CRM system has been constructed. The CRM intelligent system has a vast amount of customer data that is analyzed using data mining techniques based on the attributes of customer relationships. The procedure of implementing DM in CRM has been proposed. The ID3 algorithm's drawbacks and the enhanced algorithm's merits are examined. K-means clustering analysis and an improved ID3 decision tree are used to explain to the client. It can offer valuable customer data, cut costs for improved customer care, and increase customer happiness and the company's competitiveness in the market.
Key words: Customer relationship management; Data mining; K-means algorithm; ID3 algorithm

1. Introduction

The use of contemporary information technology to track and analyze client demand is known as customer relationship management, or CRM [1]. When prompt action is taken to preserve company interests, consumer loyalty may shift in an effort to retain current clients and draw in new ones [2]. CRM is growing increasingly and more crucial as the company model shifts from "product" to "customer" [3]. The goal of customer relationship management (CRM) is to put the customer at the center of marketing as a whole. It does this by utilizing contemporary information tools to meet practical customer demands, improving customer service, developing executive skills, and increasing enterprise profit and

customer satisfaction—all of which contribute to the "win-win" concept of customer and enterprise management. An emerging management tool that aims to strengthen the bond between businesses and their clients is the customer relationship management system. System for managing customer relationships More crucial than managing customer data is using customer relationships in data mining to gain a deeper understanding of marketing structure, consumer preferences, etc. in order to modify the marketing plan, create new items, and so on. Businesses currently lack a thorough analysis of the vast amount of consumer data that has been gathered [4-5], which puts customer service at a disadvantage. Data mining must be implemented into CRM systems in order to better analyze client relationships. Businesses may enhance customer happiness, lower customer attrition, boost revenues, and truly embody the "take the customer as the center" service concept by using data mining to identify prospective customer demand patterns and consumer behavior models[6]. This paper studies the architecture of the CRM intelligent system, the application process of data mining, and the application of the K-means algorithm and ID3 algorithm in the CRM system.

2. Data Mining

Data mining, known as knowledge discovery in the database, is a data analysis technique that processes large amounts of data [7]. It is a complex process of extracting implicit, unknown, potential but valuable information and knowledge from a large number of noisy, fuzzy, incomplete random data. Statistics, machine learning, pattern recognition, database and data warehouse theory, information retrieval theory, visualization theory, mathematical algorithms, and high performance computer knowledge and technology are all applied in data mining technologies. Neural networks, genetic algorithms, fuzzy sets, overlaying positive rejection counterexamples, statistical analysis, and rough sets are some of the data mining techniques. The primary technology for database knowledge discovery is data mining. It looks for knowledge embedded in data from a vast amount of data using particular knowledge discovery techniques.

The technique of extracting useful information from vast volumes of data using machine learning is called data mining. It has the ability to extract unknown and useful information from the data that lies underneath the original data. Database technology, artificial intelligence technology, and statistics are all combined in data mining. Additionally, it is among the most cutting-edge areas of database and decision support research. Whereas machine learning is on creating novel techniques for knowledge extraction from data, database technology concentrates on R's effective techniques for data processing and storage. Data mining uses database technology to process data at the front end. The machine learning method is used to extract useful knowledge from the processed data. Machine learning, pattern recognition, statistics, intelligent databases, knowledge acquisition, data visualization, high performance computing, expert systems, and other fields are all related to data mining, which is actually an interdisciplinary field. The knowledge from the database can be used in many fields, such as information management, process control, scientific research, decision support, etc. Different data mining tasks require different techniques, methods and means.

Generally, data mining methods and techniques can be generalized into the following main types：

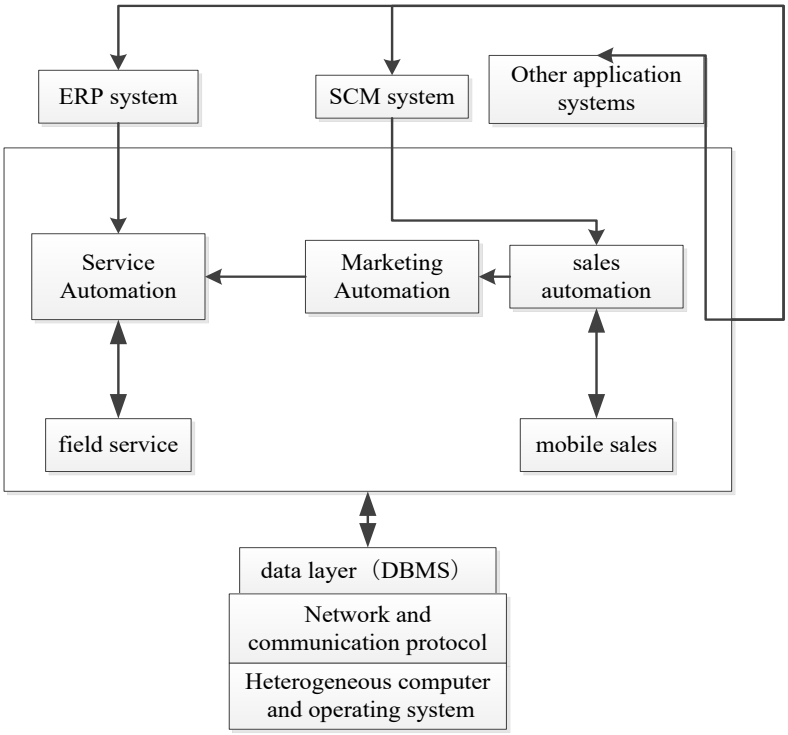
1．Statistical analysis uses traditional mathematical statistics and other mathematical principles to analyze data. It is mainly used to complete summary knowledge and related knowledge mining. Statistical analysis of attributes in the relational table is to find the relationship between them. Common methods are：

(1) Traditional subjective guidance systems: This is a system for professional applications. Analysis of financial markets is basic R technical analysis. The method is analysed from simple trend analysis to fractal theory and spectral analysis based on advanced mathematics. This technique requires empirical models as a prerequisite. They belong to the United States Metastak, SuperCharts. Candlestick, Forecaster and Wall, Street, Money.

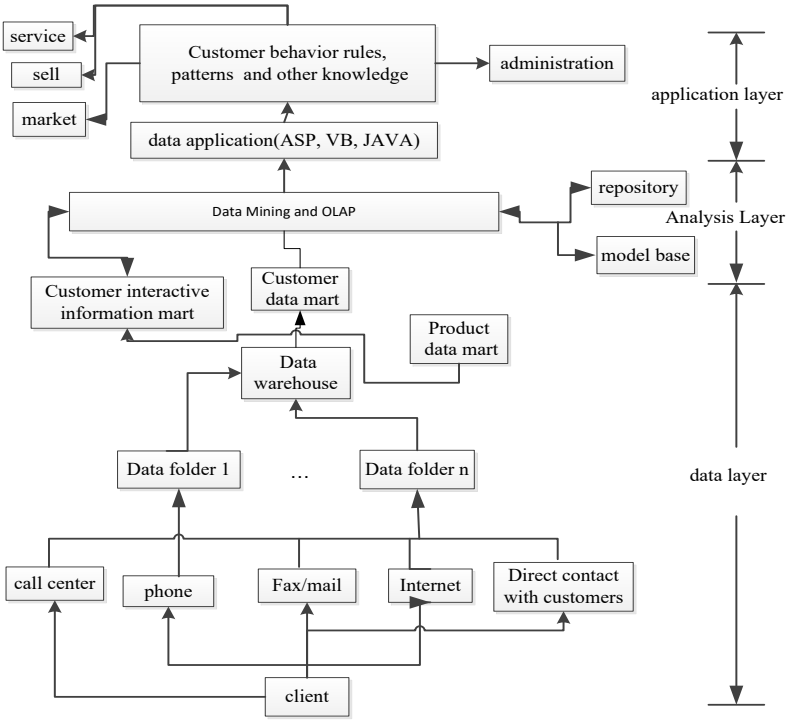
(2) Conventional statistical methods: these include factor analysis, regression analysis, and correlation analysis. Users generally assume the premise, and then use the system to verify the data. The disadvantage is that it needs to be trained to make the j, and in the process of data discovery, H1 users need to repeat a series of operations. Such items are American SAS, SPSS and Stargraphis. By Ding, in recent years the emergence and use of more advanced DM methods, these manufacturers in the original system to integrate some of the DM components, in order to achieve more perfect function.

3. The Architecture of CRM Intelligent System Based on Data Mining

Customer relationship management is "customer-centric", which enables enterprises to move out of the "product-centered" management philosophy, which is an enterprise business model. Customer relationship management to customer demand as the tenet of enterprise development, this concept will promote enterprises to maximize its related information with customers, achieve efficient service and three-dimensional management enterprise, to maximize the overall interests of enterprises. Traditional CRM systems generally include market management, sales management, product and delivery management, service support management, call center and other functions [8]. CRM, enterprise resource planning (ERP), supply chain management system (SCM), and other application systems can be integrated in order to make it easy for enterprises to make quick and effective decisions. ERP, SCM, etc. can be used as a data source CRM data warehouse, CRM market analysis forecast results can be timely feedback to the ERP system, ERP system based on the prediction of the market in a timely manner to adjust the product production and marketing strategies. The link between ERP, SCM and CRM enables the production, sales and other departments of the enterprise to respond quickly and quickly according to market forecasts. For customers is the key to finding the customer information and the relationship between the behavior patterns, applying data mining to customer relationship management system, can timely find the information such as customer demand and customer loyalty, companies can timely adjust the enterprise production and marketing strategies. Details shows in figure1.



(a) System interface



(b) Interface interactive content

Figure.1 CRM intelligent system architecture based on data mining: (a) System interface,
(b) Interface interactive content

From figure 1 intelligent CRM system based on data mining system architecture can be seen that, the data layer is the physical basis of customer relationship management system and the source of customer data, it includes the enterprise internal and external information platform and the database system. Data in different data sources is consolidated and stored in a data warehouse. A data mart is a logical subset of central data information, also known as "department-level data warehousing". Data mart data according to the business organization, a data mart contains a specific business analysis of field data, such as customer interaction information market, customer data mart and product data mart, etc. Analysis layer includes the algorithm of data mining, tools and models, etc., through the data mining algorithm of the discovery of knowledge, it is the key to intelligent customer relationship management system, through the analysis to the knowledge of the customer information, knowledge mining, knowledge classification and modeling, and other functions, to realize the change of the customer information to the customer knowledge, provide the basis for customer relationship management (CRM).The application layer includes the concept, strategy and method of customer relationship management and enterprise management decision management. The knowledge of customer behavior and pattern acquired through data mining must be quickly fed back to the management of the enterprise to improve the strategic decision-making ability of enterprise managers.

4. Data Mining in Customer Relationship Management System Implementation Process

The application of data mining in the CRM intelligent system is to extract data from data warehouse and use these data as model samples for research. Then the data mining algorithm is used to excavate the data, analyze and model the results of the excavation, and optimize the model. Specific include: the preparation and screening of the definition data of business problems; Data cleaning and data preprocessing, data mining model, evaluation and validation of the model, the construction of a practice to guide the enterprise, the feedback of customer and market information can be timely into the people into a data warehouse, quick response to enterprises; The implementation flow of data mining in customer relationship management system is shown in figure 2.

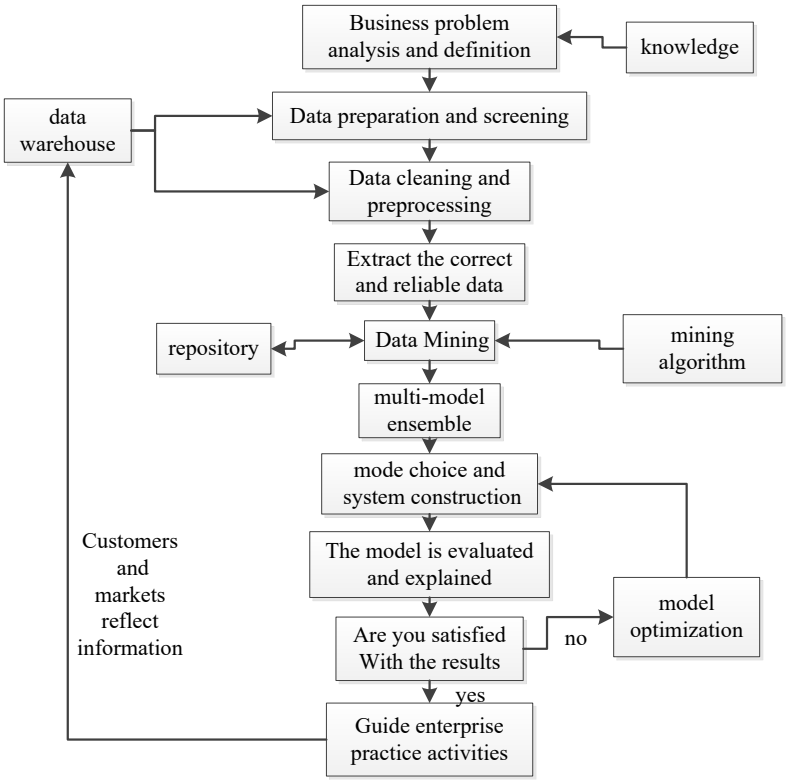


Figure. 2 Data mining flow chart

5. Applied Data Mining in CRM System

Conducting in-depth study and determining the worth of customers using data mining techniques for customer information in the customer relationship management (CRM) system [9], such as analysis of customer segmentation, to determine customer satisfaction, customer's credit rating cross-selling, analyze the market information, analysis customer turnover, customer profitability and customer value prediction, etc. The segmentation market can provide different services to different customers, and can provide targeted services to different customers, thus reducing service costs and enhancing the competitiveness of commercial enterprises. Through the user segment market, can use different marketing strategy, can strengthen customer loyalty, can through the deep analysis of customer's sales action, correctly judge the trend of the overall development of the enterprise, be effective to promote products [10]. A bank to conduct a precious metal products such as brand promotion and introduction of the activity, it can be according to the pare to principle, identify the target market customer groups, according to the result of customer identification, targeted to a few customers push marketing messages, and use the way such as a phone call to inform a few customers, both to save communication cost, and can save manpower, at the same time, Banks can also get higher profit margin.

Previous customer value analysis is simply used clustering analysis to carry out more, although these clustering methods can be the value of customers into different groups, but the lack of judgment standard, regulation of different kinds

of customer value cannot be explained well, requires policy makers to judge. Therefore, the subjectivity is very strong, different decision makers will make different decisions. There are many methods of customer value analysis. In order to reduce the subjective intention of decision makers, cluster analysis and classification analysis can be adopted. This paper mainly uses K-means algorithm and ID3 decision tree to subdivide customers. By using K-means algorithm to cluster analysis of customer value of customer samples, we can reasonably divide the collection of customer records so as to determine the category of each record. The improved ID3 decision tree can be used to explain the clustering results of customers, and find out the distribution rules of the value of each group of customers, so as to make the analysis results more meaningful.

5.1 K-means Algorithm Cluster Analysis

Based on the average distance, the customer is clustered using the K-means method. It measures the distance, absolute distance, average distance of the class and the distance from the distance. Since each customer can only be assigned to a cluster, the method is called hard clustering. The processing procedure of K-means algorithm is as follows:

The data set D has n samples. Each sample has m dimensions, and the n sample m dimensions are clustered. N samples are denoted as:

$$X=\{X_1,\ X_2,\ ...,\ X_n\}$$

Where X_i can be expressed as:

$$X_i=\{X_{i1},\ X_{i2},\ ...,\ X_{im}\}$$

Assemble into k classifications:

$$C=\{C_1,\ C_2,\ ...,\ C_k\}$$

Cluster center is $\overline{Z_i}$

$$\overline{Z_i}=\frac{1}{n_i}\sum_{J_i=1}^{n_i}P_j, i=1,2,...k$$

Randomly selected in the data set D k sample points, the sample points as a cluster of initial average, for the rest of the m - k sample points, count them and each cluster center distance, gives the value to the nearest cluster, and then, calculation of each cluster center distance, again until convergence criterion function. The square error criterion is used in this study. The formula is as follows:

$$E=\sum_{j=1}^{k}\sum_{x_i\in C}\left|x_i-\overline{Z_i}\right|^2$$

Where E is the sum of the squared errors of all points in the data set D, k is the number of categories, C_i stands for the i cluster, Z_i stands for cluster center of the i cluster. The clustering analysis steps are as follows:

(a) In the data set D, k sample points are selected, and k sample points are assigned to the original cluster center Z_i

(b) All points $X_j(j=1,2,...,n)$ in the sample set D, Calculate the distance to the cluster center Z_i: $d_{ij}=\sqrt{\left|X_i-\overline{Z_i}\right|^2}$

(c) Find the minimum distance between $X_j(j=1,2,...,n)$ and Z_i, and put

X_j in the smallest cluster with Z_i

(d) Recompute cluster centers.

(e) According to formula (5), the square error $E(t)$ of the points used in data set D is calculated and compared with the previous error $E(t-1)$

(f) If $E(t)-E(t-1)<0$, then turn to (b), or else end of the algorithm.

In the use of K-means to cluster analysis, the clustering should not interconnect and not overlap

5.2 Explain Cluster Analysis Results Using ID3 Decision Tree

The ID3 algorithm was proposed by Quinlan in 1986 to construct the decision tree. The core idea is used the information gain value as the measurement standard for the split attribute corresponding to each node in the decision tree. By using the maximum attribute of information gain, the training sample collection is divided into the minimum amount of information needed to train the set of training samples.ID3 algorithm is described as follows:

Let PN be the training set, there's p in PN and n counterexample. For a sample set, the probability of PE in the positive example set is $\frac{p}{p+n}$

A decision tree can be viewed as a source of positive, counter-example sets, message expectations information generated by the source.

$$I(p,n)=-\frac{p}{p+n}log_2\frac{p}{p+n}-\frac{n}{p+n}log_2\frac{n}{p+n}$$

Set the value of the property A is $\{A_1,A_2...A_n\}$. Classify PN as a subset of k: $\{PN_1,PN_2...PN_k\}$

Let's say PN_i has p_i positive cases, n_i counterexample, the desired information that subtree PN_i needs is $I(p_i,n_i)$. The desired information needed for the tree of A is the weighted average of desired information for each subtree.

$$E(A)=\sum_{i=1}^{k}\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

The information gain for the classification of A tree with roots is:

$$Gain(A)=I(p,n)-E(A)$$

The ID3 algorithm selects the maximum attribute of $Gain(A)$ as the branch property of the node, and is used this principle for each node of the decision tree until the complete decision tree is established.

ID3 decision tree generation algorithm is as follows:

Import: S： The training sample set is mainly described by the discrete attribute value. Candidate-attribute: Candidate attribute set.

Output: A decision tree.

Method:

(1) Create node N

(2) IF S is in a class C THEN;

(3) Returns N as a leaf node, marked as class C;

(4) If Candidate-attribute is Null THEN.

(5) Returns N as the leaf node, marked as the common class in S; //

Decision by majority

(6) Select the attribute A with the highest information gain in candidate-attribute.

(7) Mark node N is A.

(8) FOR Each known values a_i in A, division of training samples.

(9) The branch of node N that has a condition of B.

(10) Let's say A is the branch of the training sample S.

(11) If $A = a_i$ is Null THEN.

(12) Add a leaf and mark the most common class in S.

(13) ELSE plus a return node with Generate decision tree（S, candidate_A）.

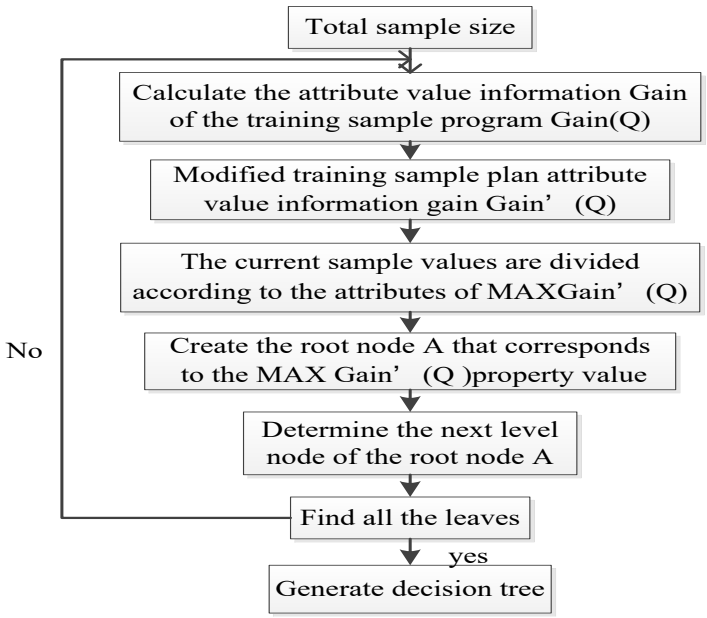


Figure.3 Steps of ID3 Algorithm

An ID3 algorithm is used to classify the data with high accuracy, low number of layers and few leaves.ID3 algorithm, however, some deficiencies: division of metrics such as ID3 algorithm using the information gain, through practice found that the proposed algorithm to choose the number of attribute value is more attribute is applicable, and the number of attribute values more is not the optimal attribute [11-12]. The ID3 algorithm is applicable to the processing of discrete attributes. For continuous attributes, the discrete processing is required, and the ID3 algorithm is not suitable for dealing with the lack of specific attribute values for an attribute. In view of the above shortcomings, the ID3 algorithm is improved and a modified parameter is introduced in the improved algorithm. After the information gain of each attribute is calculated, the parameters are calculated by using the function to modify the information gain, which is used as the division metric of attribute selection and sample partition subset. For those samples that lack attribute values, the random probability number of relative frequency is adjusted. The specific algorithm steps are shown in figure 3.

6. Conclusions

As the business model is transformed from "product" to "customer", the importance of CRM is becoming more and more important. This paper applies clustering analysis algorithm K-means and decision tree ID3 algorithm to customer relationship management system, and data mining algorithm can extract useful information from customer relationship management system. Therefore, this study designed an improved algorithm model based on decision tree ID3.Through the K-means cluster analysis and improved ID3 algorithm can extract the customer information better, provide effective information for management decisions, and improve the profits of enterprises, enterprise and customers "win-win".

The approach suggested in this study offers a new technical avenue for business customer relationship management from the standpoint of practical utility. Businesses can accomplish the following objectives by fully utilizing customer data: first, precisely determining customer needs and improving the design of products and services; second, implementing personalized marketing and increasing the marketing input-output ratio; third, anticipating trends in customer behavior and developing response plans beforehand; and lastly, allocating resources optimally and enhancing operational efficiency. Eventually, these enhancements will result in higher corporate profits and happier customers, thereby creating a "win-win" scenario for businesses and consumers.

Looking ahead, customer relationship management will become more sophisticated and individualized as big data and artificial intelligence technologies continue to advance. The approach suggested in this study offers helpful suggestions for additional research into consumer data analysis; nevertheless, in order to continuously enhance the model's performance and applicability, it must be validated on a bigger data set and taken into consideration for integration with other machine learning algorithms. At the same time, a key area for future research will be how to fully utilize data value while maintaining client privacy.

References

[1] Shu Xiao-bin, Li Min, Bank Account Manager Management System and Realization, Computer Knowledge and Technology, 2012, Vol.8, No.27,, pp.6492-6495.

[2] Sun Ning. The Research of CRM Based on Data Mining, Science Technology and Industry, June, 2011, Vol.11, No.6. pp.59-61.

[3] Zhang An-an, Xiong Hao-qi, The Research Data mining in Client Relation, JIANGXI SCENCE, 2010，Vol.28, No.5, pp.688-691.

[4] Pang Zhi-yong, The Design and Implementation of CRM for Small and Medium Enterprises, Mechanical Management And Development, 2010, Vol.25 No.5, pp.146-147.

[5] Qu Zhao, The Ideas on Structure System Modules of SMEs CRM system, Journal of Yunnan Agricultural University, 2008, vol.2, No.3, pp.10-13.

[6] Zhang Jun, Design and Implementation of College CRM System based on Web Services, INTELLIGENT COMPUTER AND APPLICATIONS, 2015, Vol.5, No.3, pp.96-101.

[7] Natchiar, S. U., & Baulkani, S. Customer relationship management classification using data mining techniques.International Conference on Science Engineering and Management Research. IEEE, 2015. Vol.37, pp.1-5.
[8] Singh A, Rana A, Ranjan J. Data mining techniques and its effect in customer relationship management[J]. International Journal of Data Analysis Techniques & Strategies, 2015, 7(4):406.
[9] Banerjee, P., & Banerjee, A. (2024). Role of data mining under e-commerce environment in understanding consumer psychology and customer relationship management. International Journal of Electronic Customer Relationship Management, 14(3-4), 318-343.
[10] Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. Neural Computing and Applications, 36(9), 4995-5005.
[11] Wang Xiao-wei, Jiang Yu-ming, Analysis and improvement of ID3 decision tree algorithm. Computer Engineering and Design, 2011, Vol.32, No.9. pp.3069-3072.
[12] Wang Xin, Zhang Yu, Building on Model of Photovoltaic Power Generation System Growth Based on System dynamics. Journal of Northeast Dianli University, Oct, 2012, Vol. 32, No.5. pp.16-19.