Hindawi
Scientific Programming
Volume 2022, Article ID 2520087, 13 pages
https://doi.org/10.1155/2022/2520087

Research Article

Research on the Application of Multimedia Entropy Method in Data Mining of Retail Business

Ting Li[1] and Can Zhang[2]

[1]*School of Logistics Management, Yunnan University of Finance and Economics, Kunming 650000, Yunnan, China*
[2]*School of Business Management, Yunnan University of Finance and Economics, Kunming 650000, Yunnan, China*

Correspondence should be addressed to Can Zhang; zz1379@ynufe.edu.cn

Received 28 December 2021; Revised 23 January 2022; Accepted 3 February 2022; Published 23 March 2022

Academic Editor: Ahmed Farouk

In recent years, database technology has developed rapidly and is changing with each passing day. With the support of network technology, its application scale, scope, and depth are constantly expanding. With the explosive growth of data, we are also faced with such a challenge: First, as a basic information storage and management method, database technology can only perform simple data processing, such as query, statistics, reports, etc.; lack of decision-making; analysis; prediction; and other advanced functions. Secondly, in the face of these massive data, people pay more attention to how to dig out the important information hidden in these data, rather than the data itself. Therefore, data mining technology, which integrates statistics, artificial intelligence, pattern recognition, and optimization, emerges as the times require. Data mining technology is application-oriented from the very beginning, and its great success in various industries has fully demonstrated its strong vitality, especially in the retail industry. If the data mining technology can be perfectly combined with the retail industry, it can not only bring great convenience to customers but also inject new vitality into enterprises, making them invincible in the fierce competition. This paper proposes a filtering of high-quality customer system framework based on maximum entropy, which expresses customer data as a feature vector for feature selection and feature smoothing. The filtering performance of different feature sets is compared by combining different characteristics of customer data. Experiment and conduct multimedia presentations. Experiments show that the filtration performance of this system is better than the general filtration system.

1. Introduction

Retail trade is an activity that involves selling goods or services directly to the final consumers. The target of this industry is more consumers than wholesalers or manufacturers [1]. This also determines that retail owners have the following characteristics: (1) As far as customers are concerned, they come to consume, and the relationship with retailers is intermittent. (2) As far as commodities are concerned, retailers adopt marketing, consignment, joint marketing, and so on. The number of suppliers connected with retailers is also very large. (3) In terms of profit, the gross margin of the retailer is lower than that of the manufacturer. (4) There is also an important characteristic that retailing is a kind of sales behavior, which is not only affected by seasons, holidays, and other external factors but also caused by its own promotion, price reduction, and so on. The so-called data mining, in business applications, is a way to analyze the huge amount of data stored in enterprises through mathematical models, to find out different customers or market segments, and to analyze consumers' preferences and behavior. The tasks of data mining include association analysis [2], clustering analysis, classification, prediction, temporal pattern, and deviation analysis: (1) The purpose of association analysis is to find out the association between data. Generally, two thresholds of support and credibility are used to measure the relevance of association rules. (2) Clustering is to classify data into several categories according to similarity. Data in the same category are similar to each other, and data in different categories are quite

different from each other [3–6]. (3) Classification is to find a conceptual description of a category, which represents the overall information of such data, that is, the connotation description of the class, and use this description to construct a model. Generally, rule or decision tree pattern is used to represent [7]. (4) Prediction is to use historical data to find out the law of change, establish a model, and then predict the types and characteristics of future data. (5) Time series pattern refers to the pattern with high repetition probability which is searched by time series. (6) There are many abnormal data in the database. It is very important to find the abnormal data in the database. It may correspond to the abnormal phenomenon in business [8–10].

There are many kinds of data mining methods. The main data mining methods used in customer relationship management of retail industry are prediction analysis, correlation analysis, and clustering analysis. Predictive analysis is generally based on the operation of the predictive analysis model designed to achieve. Predictive analysis models usually assume that some phenomena (dependent variables) arise from the appearance of other phenomena (independent variables), or change with the change of other phenomena. There is a stable quantitative relationship between dependent variables and independent variables. In this way, the possible situation can be predicted by known data [11–13]. In data mining, the construction of the prediction analysis model is usually to detect the customer's response to a particular marketing activity and the degree of reflection. Data mining technologies that can predict and analyze include logistic regression, decision tree, and so on. Logical regression is used to construct the quantitative relationship between the target variable (dependent variable) and more than one predictive variable (independent variable) [14–16]. Formally, logistic regression is very similar to linear regression. The main difference is that the dependent variables in logistic regression are not continuous variables, but discrete or categorical variables [17]. Generally, logistic regression can be used to predict two or more levels of results. But in retail business, two levels of results are commonly used, such as customer response or nonresponse to a promotional activity [18–21]. Decision tree is also used to construct the quantitative relationship between target variables (dependent variables) and more than one predictive variable (independent variables), so as to detect the attributes of target objects (such as customers or products) in a dependent variable. The method of decision tree is to divide the data of the target object according to the order of independent variables, and divide all the target objects into different groups. There is a great heterogeneity among the groups, and there is a great homogeneity within the groups [22]. Then, we find out the relationship between each factor and the target event, and use it to predict the customer's behavior [23–25]. For example, in order to reasonably classify customers, the membership level of customers is determined according to their annual consumption points. On this basis, the decision tree method in data mining technology is used to find the judgment rules to measure the valuable members. The rules can be used as the basis for judging the value of new customers and potential customers, and lay the foundation for enterprises to make targeted marketing to customers, so as to achieve win-win situation between enterprises and customers on the basis of improving customer satisfaction [26–28].

But the process of data mining is actually the process of analyzing the certainty of a data. Uncertainty analysis methods include fuzzy set theory, rough set theory, statistical entropy, and so on. In this paper, the maximum entropy model is used to verify the advantages of the maximum entropy model, which can be used to express various features conveniently, and there is no need to assume independence between features [29, 30]. The multimedia display function is very rich. The main purpose of the multimedia display is to disseminate information and promote some product content. The content to be promoted is displayed in a multimedia presentation mode, so that people can receive the information in time. This article will show the data mining on the multimedia platform and achieve good results.

2. Proposed Method

2.1. Maximum Entropy Model. In the nineteenth century, scientists proposed the laws of thermodynamics in order to study the efficiency of steam engines, and in the process of continuous in-depth research, they proposed the concept of entropy. In physics, entropy is the ratio of heat energy to temperature, indicating the degree to which heat is converted into work. In thermodynamics, entropy can represent the state of matter and the degree of chaos in the system. The greater the entropy, the greater the degree of chaos. "Information entropy" was first proposed by Shannon to realize the quantitative measurement of information. When people learn more about a random event, the uncertainty about that random event decreases. In probability theory, information entropy is actually the expectation of the amount of information. The greater the entropy, the greater the uncertainty of the event, and the entropy value of the event is determined to be 0.

Information entropy is a measure of uncertainty of random variables. The greater the uncertainty of random variables, the greater the information entropy; if the random variables degenerate to a fixed value, the information entropy is 0. For a random variable with N possible outcomes, from an information perspective, the more the information obtained, the more the uncertainty eliminated.

Entropy can be divided into individual entropy, joint entropy, and conditional entropy. In single entropy, as the name implies, the size of the event entropy is determined by a random variable. The entropy value of joint entropy is jointly determined by two random variables X and Y. Conditional entropy is the calculation of the entropy value when some information Y is known.

It can be known from the concept of entropy that the greater the entropy, the more chaotic the system and the more uniform its probability distribution. Therefore, according to the principle of maximum entropy, in the probability distribution set that meets the known conditions, we choose the optimal probability distribution as the final prediction result with the maximum entropy as the criterion.

In 1957, based on information entropy, Jaynes proposed the maximum entropy principle for the first time. He believed that a feasible solution with the largest degree of confusion (i.e. maximum entropy) should be selected from all feasible solutions. That is to say, in the process of processing information, only objective and completely definite information is added, and no artificial assumptions are added, so that the maximum entropy of the results can be obtained and all possible situations can be included. Maximum entropy principle is the criterion for choosing the statistical characteristics of random variables which are most suitable for objective conditions. Albert Einstein once said that the theory of entropy is the first law of the whole science. In nature, different random phenomena may follow the same probability distribution, and any random phenomena often follow the common probability distribution in probability theory. These common probability distributions should follow the maximum entropy principle, so the maximum entropy principle can be used as a criterion to determine the probability distribution of random variables. That is to say, using the maximum entropy formula as the objective function and combining with different constraints, we can deduce the common probability distribution in probability theory.

The establishment of maximum entropy model first needs to determine all kinds of uncertainties that may occur in the system. Then, a mathematical model with maximum entropy as the objective function and probability of occurrence of various states as independent variables is deduced to obtain the probability of occurrence of each kind of uncertainties under the condition of maximum entropy. For retail business filtering, the condition of resource allocation is to satisfy the business needs, so first we must determine the probability of high-quality customers, and then establish a mathematical model with maximum entropy as the objective function by introducing the gravity model. After the establishment of the model and the determination of the relevant parameters, the conclusions can be checked according to the examples, so as to determine that the conclusions obtained in this paper have universal applicability.

The purpose of system modeling is to construct a stochastic model to ultimately predict the stochastic process. So, the establishment of the model needs to solve two key problems. The first problem is feature selection. The selected statistics can correspond to the stochastic process of the target. The second problem is how to construct a precise model after specifying the statistics. Maximum entropy model provides a unified method to solve these two problems. Given the training data set, our goal is to select the best classification model based on the maximum entropy principle, that is, for any given input $x \in$ Input, we can output $y \in$ Output with probability $P(y|x)$. Figure 1 is a general form of the maximum entropy model framework.

In the training process of the model, we first select the features according to the data of the target training set, output the training sample set of the part, and then use the model selection algorithm to train the model. During the execution of the model, the system chooses the features of the data to be processed to get the sample, then calculates the final probability $p(y|x')$ through the model, and finally carries on the next operation.

Entropy and maximum entropy models have very rich applications in practice:

(1) Application in real life. Information entropy has important guiding significance in hydrological sequence analysis, station network layout evaluation, hydrological forecasting, water quality evaluation, water resource evaluation, and so on. In addition to the application of information entropy in the natural environment, it can also improve the quality of teaching. In order to improve the quality of teaching and guide teachers to improve their teaching level, it is necessary to evaluate the teaching quality of teachers. However, it is difficult to measure the quality of teaching. Using information entropy can comprehensively and reasonably analyze the quality of teaching evaluation, which has a very important guiding significance.

(2) Financial economy. Security risk is generally measured by the variance of security returns, but the computational complexity of variance is high; there is a problem of overestimating risk, and there is a limitation of assuming that the return distribution is normal. In order to reduce the investment risk, three investment portfolio models are established based on the information entropy, which improves the risk control in the investment process. In a complex investment environment, the concept of entropy plays a very important role in the application of different investment models, risk estimation, and rational decision-making.

2.2. Feature Selection Algorithm. Feature selection technology is an important method of data dimensionality reduction. Its essence is to select a set of optimal feature subsets that meet certain evaluation criteria from the original feature set of the original data, so as to perform classification or regression. When performing tasks, better models can be obtained and more accurate analysis results can be obtained.

Like variables, attributes, etc., features are also an aspect of data, which can be discrete data, continuous data, or Boolean data. In common classification problems, features can be divided into three categories: relevant features, which affect the classification results to a large extent and cannot be replaced; irrelevant features, which have strong random values and have no effect on the classification results; redundant features do not affect classification results or features that are associated with other features. The task of feature selection is to remove useless or redundant features from the input data, and obtain the optimal feature subset composed of related features that are most valuable for classification.

Feature Selection is also called Feature Subset Selection (FSS), or Attribute Selection. It refers to the process of

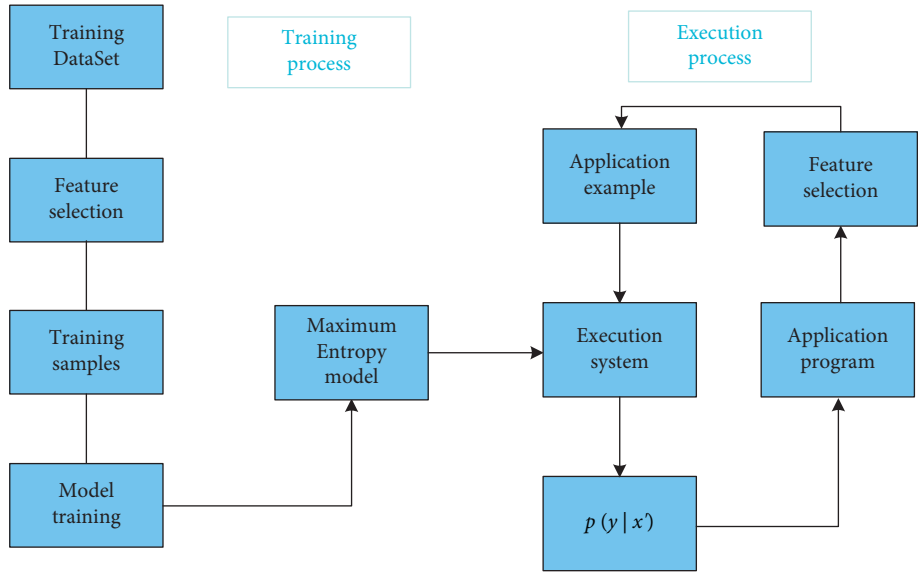

Figure 1: General form of maximum entropy model framework.

selecting N features from the existing M features to optimize the specific indicators of the system. It is a process of selecting some of the most effective features from the original features to reduce the dimensionality of data sets. It is an important means to improve the performance of learning algorithms, and also a key data preprocessing step in pattern recognition. For a learning algorithm, a good learning sample is the key to the training model. In addition, feature selection and feature extraction need to be distinguished. Feature extraction refers to the calculation of a more abstract feature set by using existing features, and also refers to the algorithm of calculating a feature.

According to the feature selection framework, the general feature selection algorithm includes four steps: subset generation, subset evaluation, stopping condition, and subset verification. Subset generation is a continuous search process. Based on the original feature set, a starting feature set is selected, and a specific search strategy is used to generate a feature subset for the next evaluation according to a certain search direction. The subset evaluation is to evaluate the feature subset generated during the subset generation process through some evaluation criteria to determine whether it is the optimal feature subset, and if so, replace the current optimal feature subset. The stop condition is set to prevent the search process from entering an infinite loop, which is generally a threshold for the number of searches or the number of features. Subset verification is the last step of the feature selection algorithm. Usually, a classifier is used to train and test the original feature set and the selected optimal feature subset to compare the pros and cons of the selected optimal feature subset.

The feature selection process generally includes four parts: generation process, evaluation function, stop criterion, and verification process. Generally speaking, feature selection can be regarded as a search optimization problem. For feature sets of size n, the search space is composed of $2n - 1$ possible states. Davies et al. proved that the search for the minimum feature subset is a NP problem, that is, besides an exhaustive search, it cannot guarantee to find the optimal solution. However, in practical applications, when the number of features is large, an exhaustive search cannot be applied because of too much computation, so people are committed to using heuristic search algorithm to find suboptimal solutions. General feature selection algorithms must determine the following four elements: (1) search starting point and direction; (2) search strategy; (3) feature evaluation function; and (4) stop criteria. The search starting point is the state point where the algorithm starts to search, and the search direction refers to the order in which the feature subset of the evaluation is generated. The starting point and direction of the search are related, and they decide the search strategy together. Generally, according to different search starting points and directions, there are four situations as follows: (1) The forward search starting point is an empty set S. According to a certain evaluation criterion, as the search progresses, the best feature is selected from the feature set not included in S to continuously join S. (2) The backward search starting point is the complete set S, and the least important features are continuously removed from the S according to certain evaluation criteria until a certain stopping criterion is reached. (3) The two-way search starts from both directions. When the middle of the feature subset space is generally searched, the subset that needs to be evaluated will increase dramatically. When using one-way search, if the search passes through the middle of the subset space, it will consume a lot of search time, so two-way search is a more common search method. (4) Random search starts from any starting point, and it has certain randomness to add and delete features [31].

Assuming that the original feature set has n features (also known as input variables), there is a possible subset of $2n - 1$ nonempty features. The search strategy is meant to find the optimal feature subset from the search space

containing $2n-1$ candidate solutions. Search strategies can be roughly divided into the following three categories: (1) Exhaustive search can search each feature subset. The disadvantage is that it will bring huge computational overhead, especially when the number of features is large, and the computational time is very long. Branch and Bound (BB) shortens the search time by pruning. (2) Sequence search avoids simple exhaustive search, and adds or eliminates features to the current feature subset according to a certain order in the search process, so as to obtain the optimized feature subset. Typical sequence search algorithms are forward and backward search, floating search, bidirectional search, sequence forward, and sequence backward search, etc. Sequential search algorithm is easy to implement, and its computational complexity is relatively small, but it is easy to fall into local optimum. (3) Random search begins with a random subset of candidate features, and approximates the global optimal solution step-by-step according to certain heuristic information and rules. For example, genetic algorithm, simulated annealing algorithm, particle swarm optimization algorithm, and immune algorithm.

In feature selection, a subset of feature sets that can express the statistical characteristics of the stochastic process is selected. Commonly used classification feature selection methods include document frequency, information gain, mutual information, expected cross-entropy, and so on. Feature selection filters out some features. A possible problem with this approach is that some useful information is ignored. Due to the small number of keywords and to avoid filtering out any useful information during feature selection, this paper adopts the solution provided by the maximum entropy model.

This paper mainly uses two incremental feature selection (IFS) algorithms, namely, basic algorithm and approximation algorithm, which are based on incremental feature selection and conditional maximum entropy method proposed by Berger et al. Each step of the iteration process is represented by the change of activity characteristic S. The current activity feature S determines the maximum entropy model P_s and model space $C(S)$:

$$C(S) = \{p \in P | E_p(f) = E_p(f), f \in S\},$$
$$P_s \equiv \arg\max_{p \in C(S)} H(p). \tag{1}$$

After adding a new feature $\hat{f}$ to S, new activity feature $\hat{f} \cup S$ and model $P_{\hat{f} \cup S}$ are obtained.

$$C(S \cup \hat{f}) = \left\{p = E_p(f), f \in S \cup \hat{f}\right\},$$
$$P_{S \cup \hat{f}} \equiv \arg\max_{p \in C(s \cup \hat{f})} H(p). \tag{2}$$

With the addition of $\hat{f}$, the model $(P_{S \cup \hat{f}})$ can better represent the characteristics of the training set and generate logarithmic likelihood gain $\Delta L(S, \hat{f})$ of training set data.

$$\Delta L(S, \hat{f}) \equiv L\left(P_{s \cup \hat{f}}\right) - L(P_s). \tag{3}$$

Each iteration process selects the feature $\hat{f}$ which maximizes the $\Delta L(S, \hat{f})$ value in the candidate feature space and adds it to the current active feature.

(1) The basic incremental feature selection algorithm estimates the model under the new feature set with iteration algorithm at each step, and calculates the maximum likelihood gain. The basic process is described as follows (Algorithm 1):

The key problem of the algorithm is that the computational complexity is too large. Each feature needs to call IIS algorithm for all candidate features, and the logarithmic likelihood of training set data is calculated. Obviously, this algorithm is not feasible.

(2) The approximate incremental feature selection algorithm is based on the basic algorithm to reduce the amount of computation. Assuming a new feature is added, only this new feature parameter is changed in the whole model, while other existing feature parameters remain unchanged, or the model after adding a new feature only depends on the original model and parameter α, that is:

$$P^{\alpha}_{s,f} = \frac{1}{Z_{\alpha}(x)} P_s(y \mid x) e^{\alpha f(x,y)} \text{ where } Z_{\alpha}(x) = \sum_{y} Ps(y \mid x) e^{\alpha f(x,y)}. \tag{4}$$

The formula for calculating the approximate gain is as follows:

$$G_{s,f}(\alpha) \equiv L\left(p^{\alpha}_{s,f}\right) - L(P_s) = -\sum \widetilde{p}(x) \log Z_{\alpha}(x) + \alpha E_{\widetilde{p}}(f),$$
$$\sim \triangle L(s, f) \equiv \max_{\alpha} G_{s,f}(\alpha),$$
$$\sim P_{s \cup f} \equiv \arg\max_{p^{\alpha}_{P_{s \cup f}}} G_{s,f}(\alpha). \tag{5}$$

This approximation simplifies the calculation of logarithmic likelihood gain caused by introducing new features into a one-dimensional optimization problem and greatly reduces the computational complexity. But, at the same time, it may lead to a problem: it is possible to choose the characteristic f with the maximum approximate gain $\sim\triangle L(s, f)$, while ignoring the characteristic $\hat{f}$ with the maximum gain $\sim\triangle L(s, \hat{f})$. The approximation algorithm is a feasible one, but it is time-consuming. On this basis, the IFS is improved, and a selective gain calculation algorithm is proposed. Each step only calculates the gain of some features which gain more in the last feature selection process, and calculates the gain of the current feature selection step.

At present, the feature selection technology is widely used in the fields of Web document processing, image processing, network security, and medical diagnosis and analysis. The research on feature selection algorithms is also more in-depth, and a large number of new algorithms are emerging. The selection of algorithms has become a very important issue. The problem. Generally speaking, in addition to considering the specific scenarios of the application, the selection of the

(1) Initialize $S = \varnothing$;
(2) For each candidate feature f, the following steps are performed:
(a) $P_{S \cup f}$ is calculated by IIS algorithm;
(b) Calculate the gain $\Delta L(S, f)$ when adding f;
(3) Check the algorithm end condition, and if the condition is true, it ends;
(4) The characteristic $\hat{f}$ with maximum gain $\Delta L(S, f)$ is selected and added to S.
(5) The P_s is calculated by IIS algorithm and go to step 2;

Algorithm 1: Basic incremental feature selection algorithm.

feature selection algorithm also needs to pay attention to the following factors: First, the scale of the data. For small-scale data sets, you can use the Filter method or Wrapper method that is close to full search, such as the BB algorithm; and when the scale of the data set is large, you should use a more efficient Filter method, such as Relief or ReliefF algorithm, etc. Second, the type of data to be processed. Different feature selection algorithms can handle different types of data. For example, the BB algorithm cannot handle discrete data. The Relief and ReliefF algorithms can handle both discrete and continuous data. MIFS and MRMR algorithms are processing for continuous data; it needs to be discretized first. Third, the category of data to be processed. For data samples whose classes are unknown, unsupervised methods should be used. Fourth, the requirements for classifier performance. If the requirements for the output accuracy of the classifier are very high, the Wrapper method based on heuristic search or genetic algorithm can be selected.

2.3. Customer Feature Smoothing. Smoothing assigns a small number of probabilities to events that do not occur. When there are enough training data, smoothing has less effect. The feature vectors transformed from customer information are sparse. For the maximum entropy of the maximum likelihood model, which is essentially an exponential form, when the feature vectors are sparse, the model will become worse. Therefore, smoothing optimization is needed to reduce or overcome the influence of over-adaptation in the training process. For those features that do not appear in the training set, it is not appropriate to simply think that the probability is zero. Generally, it is necessary to smooth them. There are relatively many studies on N-gram smoothing technology, including absolute discount, linear discount, Good-Turing method, Katz regression, linear interpolation, and so on. The research usually introduces the smoothing technology of N-gram into the smoothing of maximum entropy model. Martin's absolute discount method based on Cut-Offs and absolute discount method has better smoothing effect. Stanley compares the smoothing algorithm of the maximum entropy method with the traditional N-gram smoothing algorithm, and the performance of the Gaussian prior distribution is better among the alternative smoothing methods.

The smoothing method is a forecasting method that weighs the continuously obtained actual data and the original forecast data to make the forecast result closer to the actual situation, also known as the smoothing method or the recursive correction method. The smoothing method is a specific method in the trend method or time series method.

For the situation where the actual data is close to stationary, a smoothing method can be applied to eliminate the influence of accidental factors. Expanding the above iterative formula in time, it is directly represented by the sampling value and the estimated value 1, which means that the estimated value at time $t+1$ is the weighted smoothing of the actual sampling value in the past, and the relationship between its weighting coefficient and time. It conforms to the exponential law, so that the situation at the earlier time has less influence on the forecast. Therefore, this smoothing method is also called exponential smoothing. The value of the smoothing parameter a should be selected according to practical application experience. The larger the a, the greater the influence of the recent actual sampling value. Sometimes, in order to obtain a better correction effect, the value of a can be adjusted at any time to make it time-varying.

Absolute discount smoothing technology refers to discounting the observed events in the model, subtracting a fixed value d, and then apportioning the probability of discount to all the events that do not occur. That is, if the number of occurrences of event w is r, the probability of using absolute discount w is:

$$p(w) = \begin{cases} \dfrac{r-d}{N}, & r > 0, \\ \dfrac{(B - N_0)d}{N_0 N}, & r \le 0. \end{cases} \tag{6}$$

Among them, N is the number of all events, B is the number of different events, and N_0 is the number of no events. Because customer information is used as the value of feature function in this paper, the problem of keeping probability 1 is not involved when discounting the number of feature occurrences.

3. Experiments

JOLAP (Java Online Analytical Processing) is used in the selected technology architecture. Java Community Process JSR 69 plans to create a simplified and comprehensive unified API for OLAP services and applications. The purpose of the JOLAP specification is to deploy or interact with the Java enterprise platform. It makes full use of the Common Warehouse Metamodel (CWM), an OMG standard that defines logical OLAP structures in a vendor-independent

manner. It also leverages the Meta Object Facility, XML Metadata Interchange, and the Java Metadata Interface. The JOLAP model is a UML model consisting of some related submodels. The package consists of logical groupings of models. From this perspective, JOLAP is divided into six groups: core metadata is adapted from CWM metadata definitions, which define OLAP metadata in a vendor-independent manner. Resource models define connection and connection factories, which are based on the principles of Java Connector Architecture Common Client Interface. The resource model is different from the standard JCA implementation because it includes OLAP-style interactions. The query model defines the concepts of dimension selection, boundary, cube view, and aggregation and manipulation of dimension data. The model also contains asymmetric and transactional features. The cursor model defines how to view the dimension result set returned by the query. The source model and the server-side metadata model are defined as optional packages. The source model provides support for primitive query operations; the server-side metadata model defines other metadata for deployment-oriented classes.

System software deployment includes data warehouse products: DB2 8.1; OLAP server DB2 OLAP Server 8.1; FTL tools: DataStage; middleware: IBM WebSphere; and data presentation tools: FEnet BI. Office.

4. Discussion

The retail industry provides consumers with needed goods and related services. It is the link between production and consumption, and the final channel of the circulation link. It can be said that whoever masters the retail link will master the market. The retail industry is one of the largest and most important industries in my country's national economy with a large number of employees, a huge number of enterprises, and a large proportion of sales in GDP.

For market competition, monopoly will lead to the loss of social welfare, but the existence of excessive competition will also make the basic function of competition in rational allocation of social resources and improvement of social and economic efficiency not fully exerted. The so-called effective market concentration is to make the market concentration reach such a level that the economic efficiency of the society reaches a stable and sustained high level, which is an interval concept rather than a point concept.

Data mining can be defined in terms of technology and business. From a technical point of view, data mining is the application of a series of technologies to extract interesting information and knowledge from data in large databases or data warehouses. The extracted knowledge is expressed in the form of concepts, rules, laws, and patterns; from a business point of view, data mining is a new type of business analysis and processing technology. It is a new technology for discovering and extracting information hidden in large databases or data warehouses, helping decision-makers to find potential correlations between data and discover overlooked factors. These information and factors are critical for predicting trends and decision-making behavior.

The functions of data mining include: characterizing and distinguishing data, data characterization, and data differentiation.

The commonly used algorithms of data mining technology include: set theory method of data mining, decision tree method of data mining, genetic algorithm, and neural network method of data mining. Set theory methods mainly include methods based on rough set theory, methods based on concept trees, and learning methods that cover positive examples and exclude negative examples.

The customer data set collected in this paper includes 4658 customers, of which 2322 are high-quality customers and 2336 are general customers. 2927 (80%) of them were randomly selected as the training set and 731 (20%) as the test set. All the experiments in this paper were carried out in this data set. Recall, Precision, F1 value, and Error are selected as the evaluation indexes of filtering performance. In order to facilitate comparison, a commonly used statistical filtering method, Bayes method, is introduced.

In the whole Bayes test process, although the independence between features is not true in many cases, in application, due to its simple calculation advantages, it also has a certain better filtering performance. So this method is often used as a basis for performance comparisons with other filtering methods.

Evaluation of filtering performance usually borrows relevant indicators in the field of feature classification and information retrieval. Specifically, suppose there are a total of N customers in the test set. For the convenience of description, the variables are defined as shown in Table 1.

Where $N = A + B + C + D$, the performance of the filtering system can be measured by defining the following indicators:

(1) Recall = $A/A + C \times 100\%$, which is the "check-out" rate of high-quality customers, reflects the ability of the filtration system to find quality customers. The higher the recall rate, the fewer the "quality" customers.

(2) Precision = $A/A + B \times 100\%$, which is the "checking" rate of high-quality customers, reflects the ability of the filtering system to "find the right" quality customers. The higher the correct rate, the less likely the average customer is to be judged as a quality customer.

(3) Accuracy = $A + D/N \times 100\%$, which is the "checking" rate for all customers.

(4) Error rate: $= 1 -$ Accuracy, which is the "error rate" for all customers.

(5) F value: $F = (\beta^2 + 1)$ (Prcision $\times$ Recall)$/\beta^2 \times$ Prcision + Recall $\times 100\%$. The recall rate and accuracy rate can be synthesized as an index by F value, so that F value can fully reflect the performance of the filtering system. β is the weight factor, and $\beta = 1$ is usually used in applications.

The specific test process and analysis are as follows:

Table 1: Relevant variables for filtering performance evaluation.

	Actually for high-quality customers	Actually for general customers
The filtering system is considered to be a high-quality customer	A	B
The filtering system is considered to be a general customer	C	D

(1) The effect of feature set selection on filtering performance is to investigate the effect of customer's structural characteristics on filtering performance. This section compares the filtering performance of the maximum entropy method with Bayesian method in three cases: using only the number of customers' annual consumption, using only the cumulative amount of customers' consumption, and using all customer characteristics. The results are shown in Figures 2 and 3.

In the figure, N_Bayes represents the Bayes method that uses only the characteristics of the customer's annual consumption count, and N_ME represents the maximum entropy method that uses only the characteristics of the annual consumption times. C_ Bayes denotes the Bayes method that only uses the characteristics of accumulated consumption amount, C_ ME denotes the maximum entropy method that only uses the characteristics of accumulated consumption amount, A_Bayes denotes the Bayes method that uses all customer characteristics, A_ ME denotes the maximum entropy method that uses all customer characteristics.

By analyzing the data in Figures 2 and 3, we can draw the following conclusions: Firstly, from the perspective of recall rate, under the same feature set, the recall rate of maximum entropy method is better than the Bayes method, and the maximum entropy method has greater advantages. From the point of view of accuracy, the maximum entropy method has poor results except when the annual consumption number characteristics are used alone. In other cases, the results of the maximum entropy method are not much different from those of the Bayes method. Finally, from the perspective of F1, the maximum entropy method using all features has the best filtering effect, and the error rate is the lowest in this case.

When using the same filtering method, whether the maximum entropy method or Bayesian method, using the customer's annual consumption characteristics can effectively improve the filtering performance. The filtering performance is the best when using the characteristics of the customer's annual consumption number and the customer's cumulative consumption amount, and the worst when only using the characteristics of the customer's cumulative consumption amount, which fully illustrates the difference between high-quality customer filtering and general classification. According to the results of this experiment, the experiments in the following chapters will collect all the features of customers as customer feature sets.

(2) The influence of characteristic function on filtration performance. In this section, we will compare the filtering performance of different definitions of feature functions. Taking the basic characteristics of customers and the characteristic functions in the process of consumption as binary function, word frequency function, TF-IDF value, and χ^2, we use the Bayesian method and maximum entropy method to carry out comparative experiments. The results are shown in Figures 4 and 5.

In the figure, BVBaye represents the Bayes method in the case of using binary features, WFBayes represents the Bayes method in the case of word frequency features, ×2Bayes represents the Bayes method in the case of statistics, TI-Bayes represents the Bayes method in the case of TF_IDF, BVME represents the maximum entropy method in the case of using binary features, WFME represents the maximum entropy method in the case of using word frequency features, ×2ME represents the maximum entropy method in the case of using χ^2 statistics, and TI-Bayes represents the maximum entropy method in the case of using TF-IDF.

Through the analysis of Figures 4 and 5, we can see that when the word frequency feature function is compared with the binary feature function, the filtering performance of the word frequency feature function will be improved. The main reason is that the word frequency feature function can reflect the purchasing behavior of customers more truthfully, and it is more suitable for the realization of customer filtering based on the shopping process. The error rate of the maximum entropy method is the lowest when using the word frequency feature function.

When the maximum entropy method is used, the performance of word frequency is similar to that of the χ^2 statistical feature function and TF-IDF feature function. Although some indexes of χ^2 statistics feature function and TF-IDF feature function may be better than word frequency feature function, they are much larger than word frequency function in terms of computation. In addition, from the comprehensive index F1, the maximum entropy method is also the best. Therefore, in order to reduce the computational complexity, the frequency feature function can be used to replace the χ^2 statistical feature function and TF_IDF feature function. In this case, it will not lead to serious performance degradation. The results of the above experiments are combined with the experimental results in (1).

This paper combines experimental results with multimedia, and finally displays the data on the multimedia platform. The following picture shows the login page of the multimedia system.

After logging into the system, users can directly view the data mining results of the retail business, as shown in Figures 6 and 7:

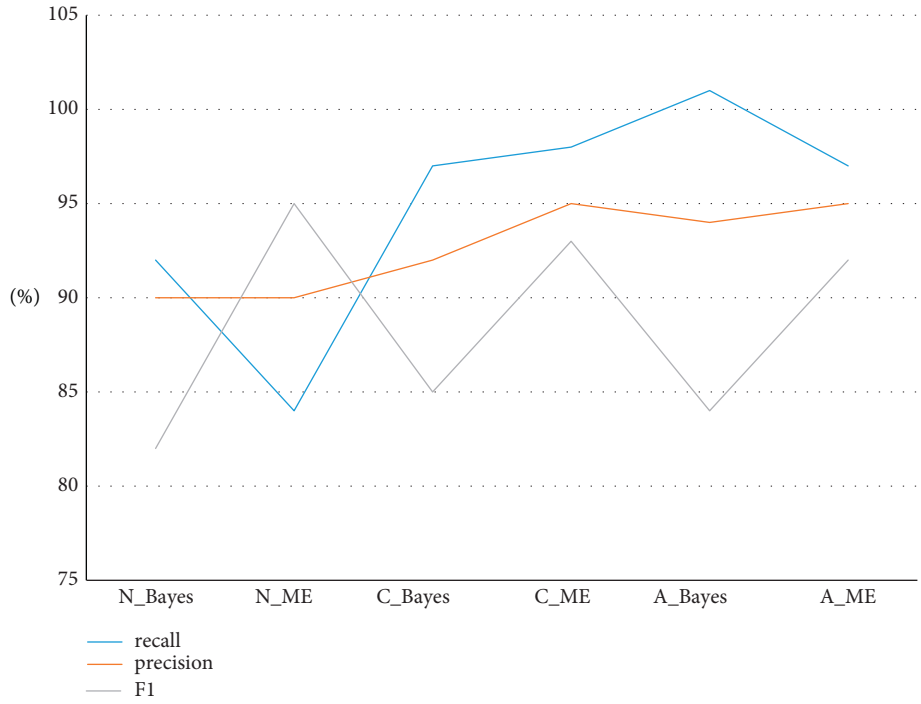

Figure 2: The impact of customer feature set selection on Recall, Precision, and F1.

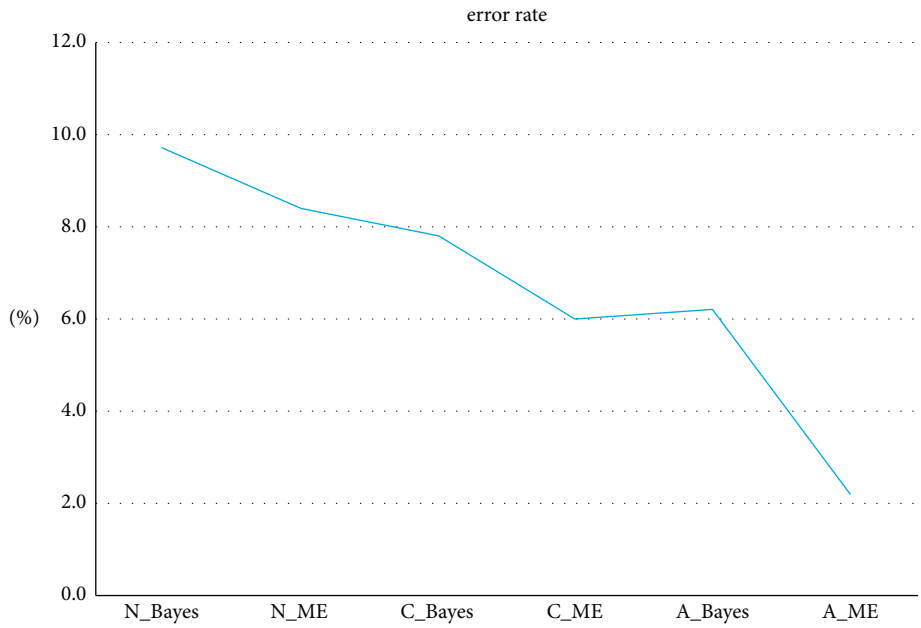

Figure 3: The impact of customer feature set selection on error rate.

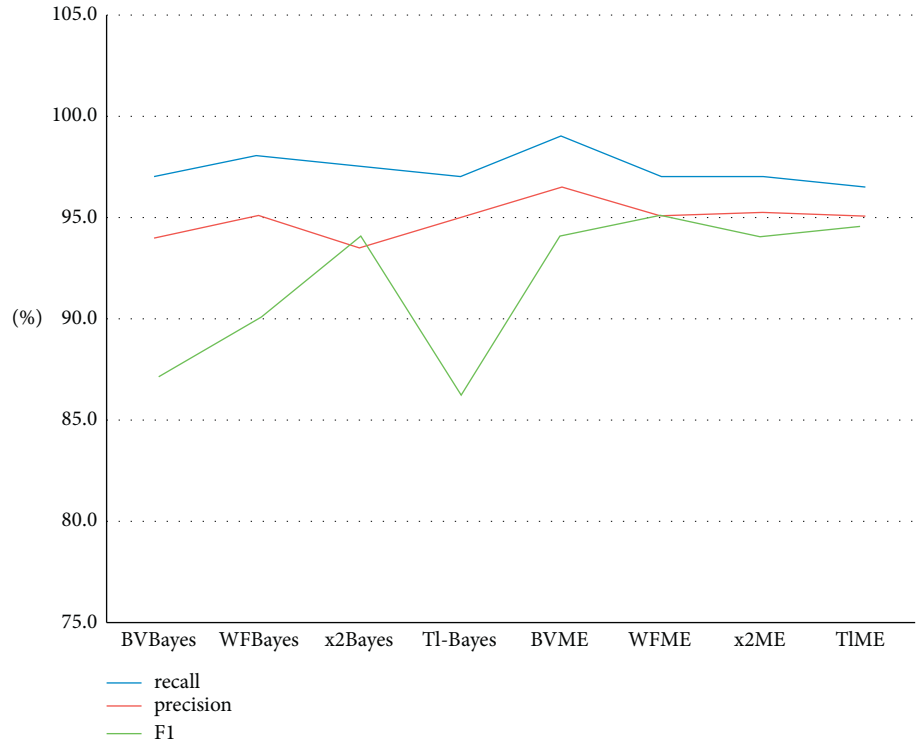

Figure 4: The influence of the selection of characteristic functions on the values of Recall, Precision, and F1.

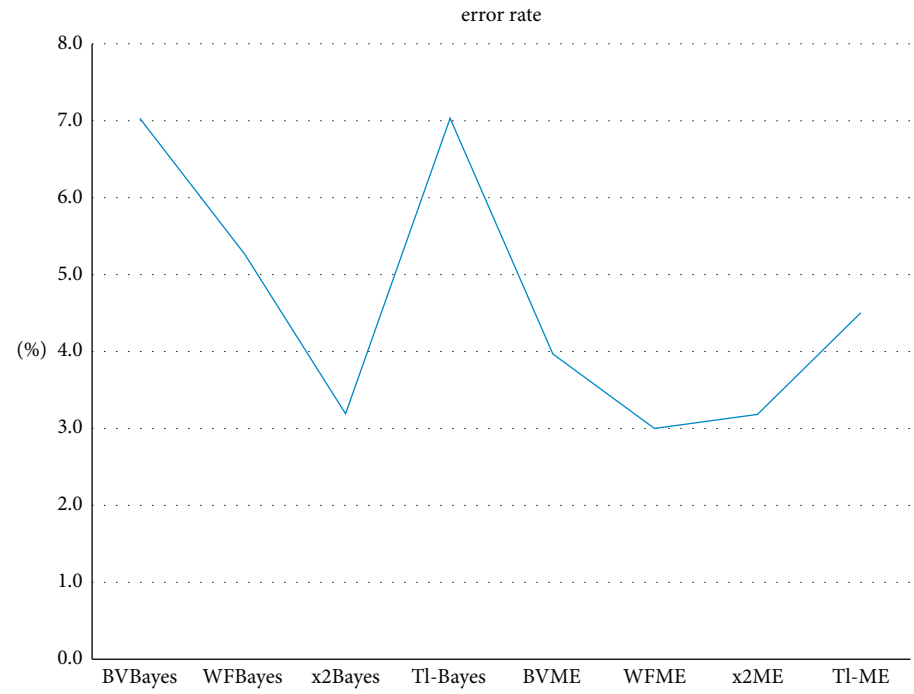

Figure 5: The influence of feature function selection on error rate.

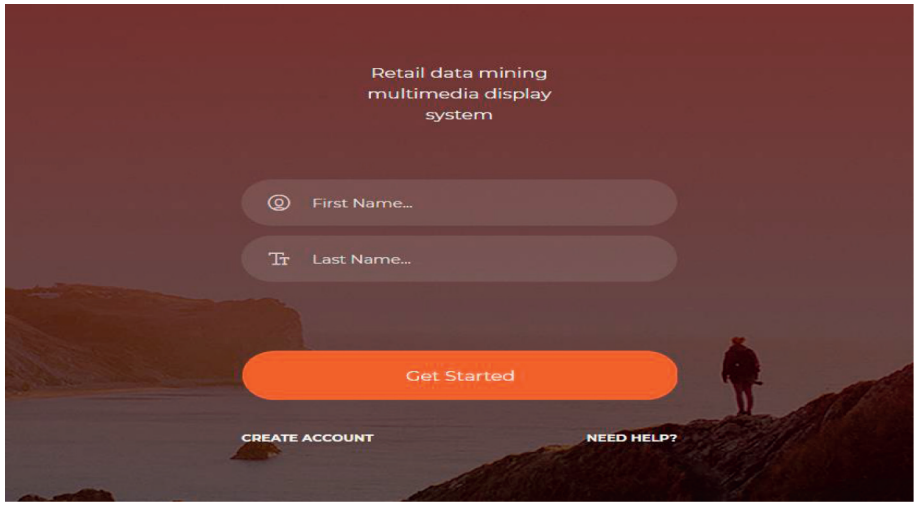

Figure 6: Multimedia system login page.

Figure 7: Multimedia data mining interface.

5. Conclusions

The application and analysis of data mining technology in the customer relationship analysis of retail industry mainly includes the design and analysis of the customer value prediction model of the retail industry, correlation analysis of customer purchase frequency, and decision tree mining model based on OLAP. Based on the analysis of existing customer data in the retail industry, a customer value prediction model is established by the decision tree algorithm to discover the different values of different types of customers, predict new customer data, and discover potential profitable customers, so as to make them become value customers that can create profits for enterprises. According to the classical *a priori* association analysis algorithm, the relationship between customers and commodity sales analysis is analyzed, and the results are applied to cross-selling or a combination marketing of commodities. Decision tree mining model based on OLAP can be used for dimension analysis and data aggregation. Decision tree mining model based on OLAP can be used for dimension analysis and data aggregation. Design of simple inference engine based on customer knowledge. The reasoning engine in this paper is a simple reasoning based on the knowledge and information obtained from the previous analysis. In this paper, the rule representation of design, the design of knowledge base, the strategy of reasoning, and the process of reasoning engine implementation are given. This paper has achieved some staged results. The results of OLAP multidimensional analysis and display based on customer data, customer value prediction model, and correlation analysis have certain value and practical significance. However, compared with some large-scale software, there are still gaps, which need to be further improved in future research. Mainly the following aspects have to be improved: (1) Because of the limitations of the mining platform, the types of algorithms available are limited. There are only two kinds of algorithms that can be used in the mining platform selected in this paper. Therefore, in future research and development, we should try to introduce new algorithms and establish new mining models. Through comparing the effects of different models, we can improve them continuously. (2) In this paper, maximum entropy is introduced into customer relationship analysis of retail industry. According to the data,

there is no formed system. Because of the short development time, the system in this paper is still in the testing stage and has not been integrated into the intelligent decision support system. In future research, the system should be integrated and perfected to realize intelligent human-computer interaction function and provide decision-makers with an intelligent reasoning engine for decision-making opinions and strategies. At present, the introduction of data mining technology into customer relationship management system research has become a hot topic for insiders. Although we choose different platforms and algorithms, we hope to find a more practical software to really meet the needs of decision-makers. The purpose of this paper was to satisfy the requirement of deep-level analysis for decision-makers, and the results have certain practicability and value.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author(s) declare no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Acknowledgments

This work was supported by the 2021 Philosophy and Social Science Planning Science Popularization Project of Yunnan Province, "Research on the Responsibility System and Performance Evaluation Mechanism for the Popularization of Social Sciences in Yunnan Province (SKPJ202154)."

References

[1] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "A tree-based statistical language model for natural language speech recognition," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 37, no. 7, 1989.

[2] L. B. Adam, F. B. Peter, A. D. P. Stephen et al., "The Candide system for machine translation," in *Proceedings of the workshop on Human Language Technology*, Plainsboro, NJ, USA, March 1994.

[3] E. Black, F. Jelinek, J. Lafferty, D. M. Magerman, and R. Mercer, "Salim Roukos, towards history-based grammars: using richer models for probabilistic parsing," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, NY, USA, February 1992.

[4] D. T. Brown, "A note on approximations to discrete probability distributions," *Information and Control*, vol. 2, no. 4, pp. 386–392, 1959.

[5] Z. Lv and H. Song, "Trust mechanism of feedback trust weight in multimedia network," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, 2021.

[6] P. F. Brown, V. J. D. Pietra, P. S. A. Della, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, 1993.

[7] P. F. Brown, J. Cocke, A. Stephen et al., "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[8] P. Brown, V. Della Pietra, P. de Souza, and R. Mercer, "Class-based N-gram models of natural language," *Proceedings, IBM Natural Language ITL*, vol. 18, pp. 283–298, 1990.

[9] P. F. Brown, P. S. A. D. Della, V. J. D. Pietra, and R. L. Mercer, "A statistical approach to sense disambiguation in machine translation," in *Proceedings of the workshop on Speech and Natural Language*, pp. 146–151, Pacific Grove, CA, USA, February 19-22, 1991.

[10] Y. Zeng, G. Chen, K. Li, Y. Zhou, X. Zhou, and K. Li, "M-skyline: taking sunk cost and alternative recommendation in consideration for skyline query on uncertain data," *Knowledge-Based Systems*, vol. 163, no. 1, pp. 204–213, 2019.

[11] M. Thomas, *Cover, Joy A. Thomas, Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 1991.

[12] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[13] Z. Lv, R. Lou, J. Li, and H Song, "Big data analytics for 6G-enabled massive internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 99, 2021.

[14] I. Csiszar, "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," *Annals of Statistics*, vol. 17, no. 3, pp. 1409–1413, 1989.

[15] L. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics & Decisions, Supplemental Issue*, vol. 1, pp. 205–237, 1984.

[16] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.

[17] S. D. P. Pietra, V. J. D. Pietra, J. Gillet, J. D. Lafferty, H. Printz, and L. Ures, Inference and estimation of a long-range trigram model,," in *Proceedings of the Second International Colloquium on Grammatical Inference and Applications*, pp. 78–92, Alicante, Spain, September 1994.

[18] X. Li, H. Liu, W. Wang, Ye Zheng, H. Lv, and Z. Lv, "Big data analysis of the internet of things in the digital twins of smart city based on deep learning," *Future Generation Computer Systems*, vol. 128, pp. 167–177, 2021.

[19] D. P. Stephen, D. P. Vincent, and J. Lafferty, *Inducing Features of Random Fields*, Carnegie Mellon University, Pittsburgh, PA, USA, 1995.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via theEMAlgorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–22, 1977.

[21] S. Guiasu and A. Shenitzer, "The principle of maximum entropy," *The Mathematical Intelligencer*, vol. 7, no. 1, 1985.

[22] E. T. Jaynes, "Probability theory as logic," in *Maximum Entropy and Bayesian Methods*, W. T. Grandy and L. H. Schick, Eds., pp. 1–16, Kluwer, St. Louis, MO, USA, 1990.

[23] C. H. Wu, Z. Yan, S. B. Tsai, W. Wang, B. Cao, and X. Li, "An empirical study on sales performance effect and pricing strategy for e-commerce: from the perspective of mobile information," *Mobile Information Systems*, vol. 2020, Article ID 7561807, 8 pages, 2020.

[24] F. Jelinek and R. L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Proceedings, Workshop on Pattern Recognition in Practice,* Amsterdam, The Netherlands, 1980.

[25] J. Lucassen and R. Mercer, "An information theoretic approach to automatic determination of phonemic baseforms," in *Proceedings of the, IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA, March 1984.

[26] B. Merialdo, "Tagging text with a probabilistic model," in *Proceedings of the, IBM Natural Language ITL*, pp. 161–172, Paris, France, 1990.

[27] A. Nádas, R. Mercer, L. Bahl et al., "Continuous speech recognition with automatically selected acoustic prototypes obtained by either bootstrapping or clustering," in *Proceedings of the, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1153–1155, Atlanta, GA, April 1981.

[28] I. S. Sokolnikoff and R. M. Redheffer, *Mathematics of Physics and Modern Engineering*, McGraw-Hill Book Company, 2nd edition, 1966.

[29] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[30] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, Philadelphia, PA, USA, October 2002.

[31] L. Adam, V. Pietra, and S. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, 2002.