**Customer purchase analysis based on KNN (k- nearest neighbours) algorithm using python.**

Akhil Sannepalli

Computer science and engineering (IOA)

Bharath University

Tambaram, India

[akhil.sannepalli2004@gmail.com](mailto:akhil.sannepalli2004@gmail.com)

Chandana Sappa

Computer science and engineering (IOA)

Bharath University

Tambaram, India

[chandusappa07@gmail.com](mailto:chandusappa07@gmail.com)

Suvan reddy Marthala Computer science and engineering (IOA) Bharath University Tambaram, India [marthalasuvanreddy@gmail.com](mailto:marthalasuvanreddy@gmail.com)

Dehith kumar Lanka

Computer science and engineering (IOA) Bharath University Tambaram, India

[lankadehithkumar@gmail.com](mailto:lankadehithkumar@gmail.com)

**Abstract**

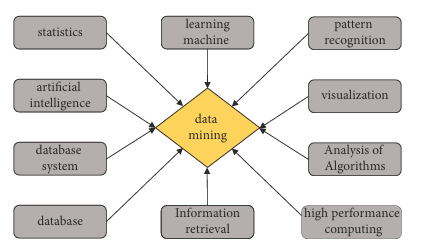
Due to today’s transition from visiting physical stores to online shopping, predicting customer behaviour in the context of e-commerce is gaining importance. It can increase customer satisfaction and sales, resulting in higher conversion rates and a competitive advantage, by facilitating a more personalized shopping process. By utilizing clickstream and supplementary customer data, models for predicting customer behaviour can be built. This study analyses machine learning models to predict a purchase, which is a relevant use case as applied by a large clothing retailer. The results indicate that a k nearest neighbour is best suited for the prediction task, showing the best performance results, reasonable latency, of fering comprehensibility and a high robustness. Regarding the different data types, models trained on sequential session data outperformed models trained on the static customer data by far. The best results were obtained when combining both datasets.

**K-nearest Neighbor**

As mentioned before, the KNN is a learning algorithm, where training data is simply stored and a test data point is awaited for classification. All of the stored training instances correspond to points in an n-dimensional feature space. A point’s nearest neighbors are defined by distance measurements, most commonly Euclidean distance. An unlabeled test data point will be assigned the label most common amongst its k-nearest neighbors. The advantages of this method are its robustness to noisy data and a very fast training speed. Disadvantages are an increased complexity of dimensionality through irrelevant features and therefore a decreased performance, highlighting the importance of feature engineering, longer prediction times compared to eager learning models and a low comprehensibility with high-dimensional input. The KNN model has been applied regularly in e-commerce regarding recommender systems, where products are recommended to a web shop visitor based on the preferences of its nearest neighbors. For classifying customer behavior on the other hand only one study could be found. The aim of the research was to classify customer web shop sessions in buying or browsing sessions. As a model, a KNN algorithm was used with clickstream data split by sessions as input.

**Research Method**

Data Mining. Market adaptability of price prediction is not strong. Many developers in order to adapt to the changing market environment, sometimes arbitrarily lower prices or raise prices, this random pricing method is not scientific. Direct price reduction not only seriously damages the interests of e-commerce agricultural products enterprises but also is easy to reduce the goodwill of the project and lose the market. The concept of data mining broadly refers to the whole process of finding valuable patterns and knowledge in a large amount of data. This process generally includes data preparation, data preprocessing, data mining model evaluation, and presentation. In a narrow sense, it only refers to a single stage of algorithm execution in the process of processing. Data mining is generally related to computer science achieves these goals through a number of methods such as statistics, online analytical processing, intelligence retrieval, machine learning, expert systems (relying on past rules of thumb), and pattern recognition. The problem of big data has been discussed by more and more people. Massive data is stored in Internet databases or other data storage devices in various fields related to our daily life, such as business, social sciences, engineering, and medical treatment. The surge of available data is the result of the rapid development of software and hardware in a highly information society, and the data collection capacity and storage devices are increasing. Generally, the pricing methods of e-commerce agricultural products marketing projects mainly include market comparison pricing method, cost pricing method, market competition pricing method, and future income pricing method. However, three factors usually need to be considered in the e-commerce agricultural products pricing practice: one is cost, including land cost, network cost, tax, and other expenses; the second is competition, the total market supply and demand, and the price of direct and indirect competitors; the third is the consumer, which is the price the target consumer can accept. therefore, the above pricing method should be used comprehensively. In the business world, for example, huge amounts of data are generated every day, including sales records, stock trans actions, product descriptions, promotions, company profits, and user feedback: Medical and health organizations have access to vast amounts of data such as medical records, patient monitoring, and medical images. Network media has become a continuous data source, such as constantly uploaded images, videos, and text information. +e data obtained in these fields is almost endless. +e wide application of these explosive growth data makes human life enter the data age. Data mining is driven by the urgent need for powerful and comprehensive tools to automatically discover and analyze this data into regular knowledge for human use. In recent years, data mining has attracted great attention in the information industry, the main reason is that 4 Computational Intelligence and Neuroscience. OA articles are governed by the applicable Creative Commons License there is a large amount of data, which can be widely used, and there is an urgent need to convert these data into useful information and knowledge. The information and knowledge acquired can be used in a wide variety of applications, including business management, production control, market analysis, engineering design, and scientific exploration. Data mining is a young and developing subject, which will certainly help us develop rapidly from the data age to the real information age. Data mining is a brand new discipline, and as an application-driven research field, data mining also belongs to an interdisciplinary discipline, database system, artificial intelligence, as shown in Figure 1. Data mining is an inter disciplinary research of great significance for the successful development and wide application of data mining. In this regard, e-commerce enterprises should combine the psychological characteristics of consumer groups, formulate scientific and reasonable price strategies, stimulate consumers’ consumption desire, drive consumption behavior, and improve economic benefits. Different consumer groups have different needs, even if the product is the same, it is very likely to have different preferences. +e reason is that the product brings different experience effects to consumers, and the degree of consumers to accept the product price and the experience effect is closely related. As a general technology, data mining can be applied to almost any field, as long as the data in these application fields has practical significance. +e most basic application areas are data stored in databases, data in data warehouses, and transactional data. It consists of a set of relational data and the software that accesses and manages that data. Software programs are responsible for defining data structure data storage concurrent and sharing distributed data access information consistent and secure authorization operations. The resources in the database can be fully managed, and a kind of control about the data can be realized the utility program of the database can make the database established on a relatively complete basis, and the database can be maintained under a relatively complete database system. Benefit is always the first driving force of marketing. The project is unsalable. Generally speaking, marketers should avoid direct price reduction and can adopt a variety of disguised price strategies according to the initial pricing to deal with the consumer psychology of buying out or not buying out. For example, we can increase the number of marketing, increase quality services, or take other measures such as commodity surcharges to attract customers and enhance customer confidence. Relational database is the most widely used database. It is a collection of data tables. Each data table contains multiple attribute columns, which can store a large number of records. Each record is identified by a unique primary key. Data stored in a relational database can be manipulated with relational query statements. Query statements can carry out specific data set operations. Given a query, query statements can be converted into connection selection projection and other relational operations to obtain user demand information. Relational operations can only carry out basic operations. We can conduct in-depth data pattern analysis through data mining technology.



Data mining visualization can efficiently and clearly show data through graphics and images, so it has a very wide range of applications. The relationship between data can be intuitively discovered through visualization analysis technology, especially in the case of small amount of data, after data visualization processing. In the process of data cleaning, noise data can be effectively eliminated. In the process of algorithm execution, especially in the process of supervised learning, the algorithm execution process can be displayed, and human-computer interaction interface can be provided. In the final knowledge display, results can be clearly dis played. Visualization techniques can be used throughout the data mining process. The process of data mining is processed according to the steps of data collection, data preprocessing, other relational operations to obtain user demand mode evaluation and knowledge display, accompanied by visual analysis technology

**Data collection**

It is the process of collecting data from various data sources, such as database systems, file systems, Internet resources, multimedia data [19]. Consumers have different willingness to pay, and some people think that the product symbolizes status, regardless, the price level will choose to buy. In this regard, e-commerce enterprises should quickly adjust their pricing strategies, make accurate judgment on consumers’ product demand and purchasing power by analyzing big data, and implement the differentiated pricing strategy to meet the individual needs of consumers and realize the maximization of the interest of enterprises within the price elasticity range.

**Data pre processing**

This includes two steps: data cleaning and data integration. Data cleaning can be used to remove noise from data and correct inconsistencies.

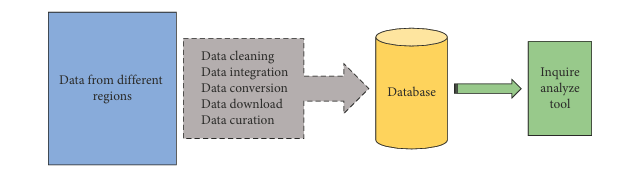
**Data integration**

It is combining data from multiple data sources into a consistent data store, such as a data warehouse.

**Data selection**

Data selection is the process of selecting the data related to the analysis task in the data warehouse system

**Computational Intelligence and Neuroscience**

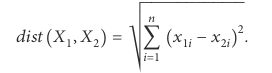
****

on the basis of users and professional knowledge, such as the implementation of data mining algorithms, the results of general algorithms are not necessarily meaning and value models. Model evaluation needs to determine the accuracy and effectiveness of the model. The final evaluation is a useful model, especially the data visualization analysis technology can be applied to show the whole process of data mining.

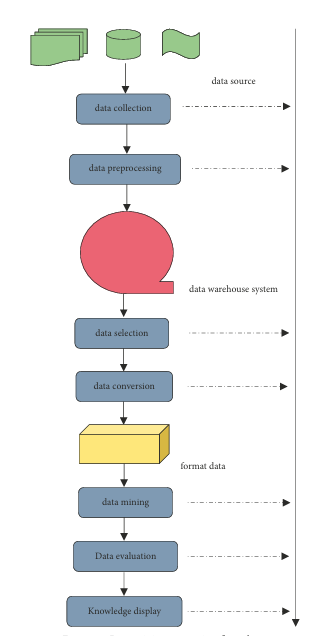
**Classical KNN Algorithm**

K-nearest neighbor algorithm (KNN), in order to distinguish it from the improved KNN algorithm, it is called the classical KNN algorithm. Classical KNN algorithm is a learning algorithm based on analogy, that is, the algorithm of prediction between prediction tuples and training tuples.Training tuples are described by attribute values, and there is a known social effect majoring, so that each training tuple is like a function of dimensional space, as follows:f(X)=f(x1,x2,……,xn).

x1,x2….,xn represent the attribute value, and f(X) represents the classification label. Query statements can carry out specific data set operations. Given a query, query statements can be converted into connection selection projection and other relational operations to obtain user demand KNN in the pattern space, and this k tuple is the K-nearest neighbor of the unknown tuple. The similarity is calculated by the Euclidean distance formula, as follows



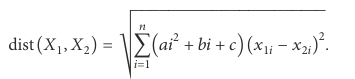
The classical KNN algorithm can be used for numerical prediction, that is,to return the true predicted value according to an unknown tuple.

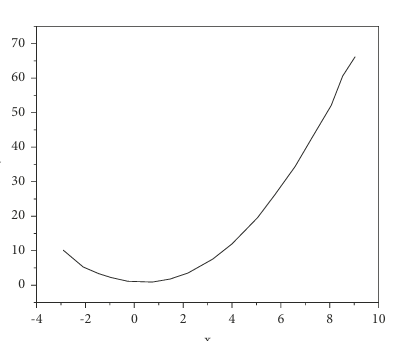


**Improved KNN Algorithm**

The classical KNN algorithm has the following two defects. First, the education formal is used to measure the development, and the same weight is assigned to each attribute value. Therefore, the prediction accuracy will be seriously affected when there is noise data. Even if weight is assigned to attribute values, the determi nation of weight is still a problem to be solved because of the multidimensional characteristics of prediction sequence. Secondly, the selection of value can only be determined through experiments. Generally, it starts from 1 and eval uates the prediction error of training tuples. As the k value changes, the value that minimizes the training error will be selected for constructing the prediction model because the time complexity is relatively high and user guidance is needed. Aiming at the two defects of classical KNN query statements can carry out social data set operation.

At this point, the attribute value of the prediction sequence (x1,x2,...xn) is regarded as the abscissa value, and the weight value is regarded as the ordinate value, and then the first improvement of the improved KNN algorithm is obtained, namely, the binomial function Euclidean distance formula, namely





PSO Algorithm Optimization of Parameters. Particle swarm optimization algorithm (PSO) is an evolutionary algorithm, an optimization algorithm proposed by American scientists in 1995 by simulating the process of achieving social behavior goals (such as group foraging and avoiding natural enemies) through the cooperation between individual behaviors of birds or fish groups.

The PSO algorithm can be well applied to the parameter optimization of BPNN algorithm and improved KNN algorithm. When PSO algorithm is used to optimize the parameters of the BPNN algorithm, the topological structure of neural network model should be defined according to the specific application of the algorithm.

the parameters to be optimized of the BPNN algorithm and the improved KNN algorithm are randomly initialized into the positions and velocities of a group of particles. Then, the fitness of each particle is calculated according to the fitness function. According to the specific application of, the fitness function is defined as

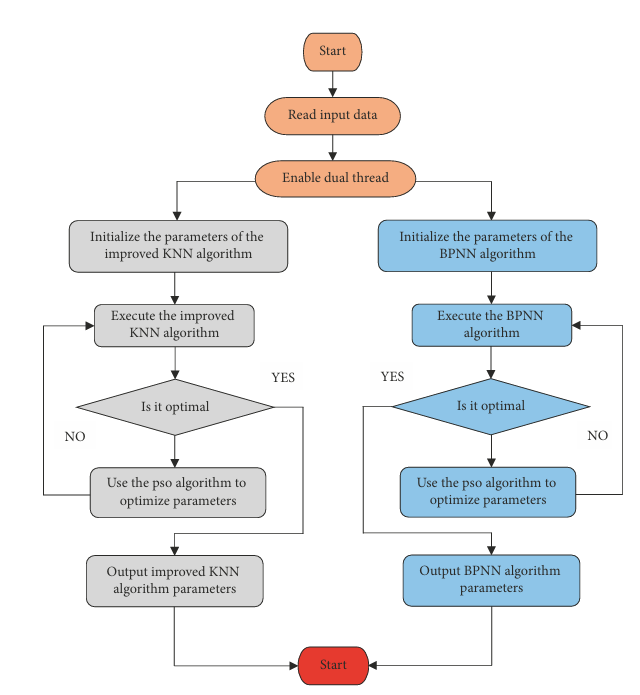


Value N is the number of training samples, T and y, respectively, represent the real value of each sample in the training sample set and the actual value of training output. Value E represents error, and the smaller the error, the higher the fitness. +e fitness function determines the local optimal position of the current particle Pbest and the global optimal position of the population Gbest, and then according to the velocity change formula.



The formula is proposed by 1998 to calculate the change speed of particles, in which ω is the inertial weight of particles, which can control the search space of particles. A large value can ensure the search in a wide space, while a small value can ensure the convergence of parameters. In the experiment, the value is 0.75, C1 and C2 represent the cognitive ability of the particle, which can control the local optimal position Pbest and Gbest, respectively. +e value 1.25 in the experiment is a random number between 0 and 1, and ran d is used to randomly assign the position 0 to 1 and speed of the particle. The velocity value can be calculated.

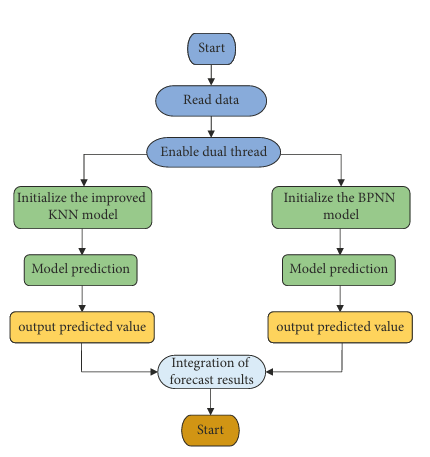
consumers will choose the commodity or combination of goods that meets the maximum utility. The training samples and Interpretation of Result



**Interpretation of Result**

Prediction Results Integration Strategy Selection. The above paper describes three strategies for the integration of prediction results. Through several experiments with test samples, the results are shown in Table 1. The prediction error of strategy 2 is the smallest, and the final prediction model chooses strategy

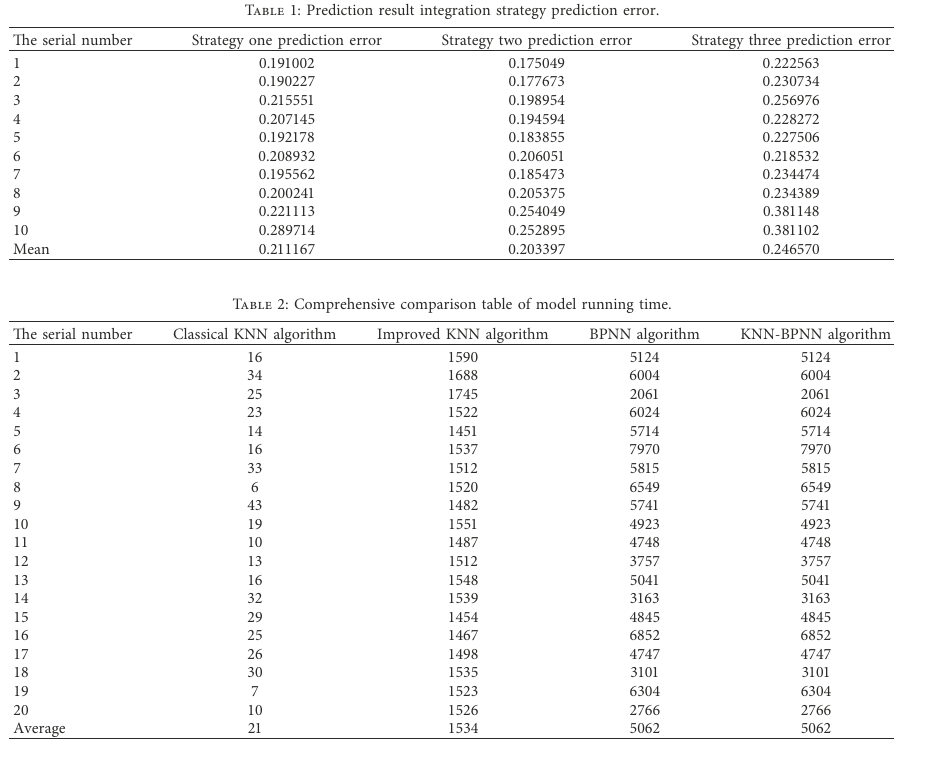
2. Price strategy is the most important part of the marketing portfolio. Compared with other marketing strategies, price strategy is the most difficult factor to determine in the controllable factors of enterprises, which requires enterprises to pay more attention to the price decision of commodities. Price strategy includes stage price strategy, geographic pricing strategy, and discount and concession strategy. Promotion refers to the promotion work carried out by enterprises in order to influence consumers’ purchasing behavior, stimulate consumers’ purchase desire, and increase the sales volume of products.

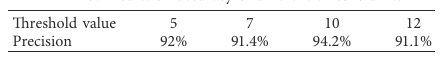


**Prediction Model Performance Comparison**

Although the classical KNN algorithm has a large prediction error, it has the strongest stability, which proves that the improved KNN algorithm has a good stability. Compared with the classical KNN algorithm, the improved KNN algorithm can not only reduce the prediction error by about 10% but also guarantee the stability of prediction, which fully proves the superiority of this algorithm. +e prediction error of the BPNN inconsistency of the quantization range is about 1% lower than that, but there is a large difference in the order of stability. Finally, the prediction model based on the improved KNN algorithm and BPNN algorithm can ensure the lowest prediction error and strong stability.

The experimental results are the classical KNN algorithm, the improved KNN algorithm, the BPNN algorithm and the actual output of the prediction model based on the KNN and BPNN algorithm and the predicted output results. The classical KNN algorithm has the maximum shadow area, and the prediction model based on the combination of the two algorithms has the minimum shadow area, that is, the minimum prediction error. As for the computational efficiency, since the database data is updated in the unit of day, it has low requirements in the specific application. However, from the perspective of algorithm, using the PSO algorithm to optimize parameters can achieve fast convergence and effectively reduce the running time.



****

intention to buy products, with the highest accuracy of 94.2%.At the same time, it also provides a good data basis for scheme design in the future precision marketing process, which can provide accurate message push to target users. The specific representation of a price strategy uses a 5-levelLikert scale(never, occasionally, sometimes, often, and always) approach.

**Conclusion**

This paper studies the prediction inconsistency of the quantitative range of the model and uses the data mining algorithm to establish the e-commerce agricultural product price and purchase intention prediction model with high prediction accuracy, strong prediction stability, and low computational complexity. In the network environment, marketing means are diverse, and marketing forms are gathered. Specific marketing methods will have a significant impact on the compromise effect, and the compromise effect shown in consumers’ purchase decisions needs to be further studied in detail. Secondly, for the alternative commodity set, whether the price difference between extreme options and discount options and the discount intensity in the price strategy will also affect the discount effect. On the basis of classical nearest neighbor algorithm, this paper combines binomial function with Euclidean distance formula when calculating nearest neighbor by similarity. Particle swarm optimization is used to optimize the parameters of binomial function coefficient and K value of nearest neighbor algorithm, and an improved nearest neighbor algorithm based on binomial function to calculate the weighted Euclidean distance is proposed. It is expected that this will play a guiding role in the production and sales of e-commerce agricultural products. The results are as follows: (1) a series of evaluation of the inconsistency of the quantization range model based on the KNN-BPNN algorithm was carried out from the accuracy stability and computational efficiency of the prediction model. By comparing with the BPNN algorithm optimized by the PSO algorithm and KNN algorithm improved by the classical KNN algorithm, the experiment proves that the inconsistency of the quantization range of model based on the KNN-BPNN algorithm has the highest prediction ac curacy. (2) Based on the excellent performance of the classical KNN algorithm in stability, the stability of in consistency of the quantization range of model established based on the KNN-BPNN algorithm also has a good performance. In terms of time efficiency, since the agricultural inconsistency of the quantization range model established based on the KNN-BPNN algorithm needs to optimize the parameters of the two algorithms, at least the time of parameter optimization of the BPNN algorithm is required even in the case of dual threads, but for the specific ap plication, its importance is relatively low. In conclusion, the above experiments prove the feasibility and effectiveness of the e-commerce agricultural product price prediction and purchase intention model based on the KNN-BPN algorithm. The overall trend of the development of data mining technology has gradually changed from dealing with simple mining problems to solving complex mining problems.

**Data Availability**

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

**Conflicts of Interest**

The authors declare no potential conflicts of interest.