SCIENTIFIC OASIS

Decision Making: Applications in Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

DECISION MAKING: APPLICATIONS IN MANAGEMENT AND ENGINEERING

Analysis and Application Research of E-Commerce Financial Management Based on T-DPC Optimization Algorithm

Yilan Wang[1], Yao Shan[2*]

[1] Accounting Department, North China Institute of Science and Technology, Langfang 065201, China
[2] School of Emergency Technology & Management, North China Institute of Science and Technology, Langfang 065201, China

ARTICLE INFO

Article history:
Received 30 August 2023
Received in revised form 20 January 2024
Accepted 31 January 2024
Available online 9 January 2024

Keywords:
Financial data; DPC; T-SNE; Prediction; Clustering.

ABSTRACT

Given the intricate, multifaceted nature of financial data in e-commerce enterprises, this article presents a T-DPC algorithm for analyzing financial management in these businesses. The algorithm utilizes the t-SNE method to reduce the dimensionality of financial data, whilst also implementing an enhanced DPC algorithm based on the K-nearest neighbor concept to analyze financial data clusters. The results show that the F-measure metrics of the DPC algorithm optimized by t-SNE improve 16.7% and 3.07% over the DPC algorithm after testing on the PID and Wine datasets, and its running time is faster than the DPC algorithm on the Aggregation, D31, and R15 datasets by 16.2. Therefore, the algorithm has reference significance for the financial analysis of e-commerce enterprises.

1. Introduction

Abbreviations

Abbreviated Name	Full Name
KNN	k-NearestNeighbor
DPC	Density Peak Clustering Algorithm
t-SNE	t-distributed Stochastic Neighbor Embedding
T-DPC	t-distributed Stochastic Neighbor Embedding-Density Peak Clustering Algorithm
ACD	Autoregressive Conditional Duration
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
AP	Affinity Propagation
Acc	Accuracy
AMI	Adjusted Mutual Information
ARI	Adjusted Rand Index

Over time, society and technology have progressed, allowing for new innovations that benefit people [1]. However, the copious amount of information and data surrounding people's daily lives has both advantages and disadvantages. While accurate information can enhance lives, erroneous information can lead to faulty reasoning [2]. Therefore, data mining is essential for conducting

[*] Corresponding author.
E-mail address: *shanyao2022@yandex.com*

https://doi.org/10.31181/dmame722024943

analysis. Cluster analysis, as an unsupervised learning method, can capture the features of each class in the data [3]. Given the recent rapid growth of the e-commerce industry and intense competition among companies, thorough financial data analysis of this sector is urgently required [4]. The DPC algorithm calculates the distance between any two points within the dataset, irrespective of high-dimensional data. This algorithm centers on clusters that are surrounded by points with low local density, furthermore, the distance between each cluster center is relatively far. Consequently, DPC exhibits two key characteristics in its simplicity and efficiency to handle clustering problems. Therefore, this study aims to conduct a financial data clustering analysis of e-commerce enterprises using the DPC algorithm. Concurrently, in view of the complex and high-dimensional characteristics of industry data, the t-SNE algorithm is used to do data dimensionality reduction and optimization processing, and use K-nearest neighbor idea to improve the artificial setting problem of truncation distance in DPC algorithm, so as to provide guiding opinions for financial analysis of this industry.

The study's significance is in categorizing and assessing e-commerce companies using the enhanced T-DPC algorithm. The algorithm employs five secondary measures; enterprise net profit, earnings per share, return on equity, operating income, and net asset per share, to divide these companies into two groups: those with good financial performance and those with bad financial performance. This categorization enables early financial warning to be given for future e-commerce enterprises. The insufficiency of this study is that only the financial data of 40 listed companies were considered for cluster analysis. To improve the study, the number of companies whose financial data is analyzed needs to be increased. Furthermore, the decision graph used to obtain the cluster centre in the K-Nearest algorithm is not sufficiently intelligent. To address this issue, it is necessary to improve the means of extracting the data cluster centre in the future. The study enhances the current literature by employing both t-SNE and K-Nearest Neighbor algorithms for optimizing the DPC clustering analysis. The former aids in dimension reduction, while the latter improves classification effectiveness, leading to the development of a proficient financial data mining algorithm.

2. Related Work

Scholars at home and abroad have conducted a wide range of research in the direction of financial analysis and financial management. Cohen *et al.,* [5] considered the characteristics of public financial management, implemented several innovations and reforms covering different areas and scopes, and proposed a future research agenda that outlines the efforts and challenges faced by public administration by analyzing the impact of institutional and general environmental factors [5]. Qiu [6] proposed a data mining approach based on financial risk management, and the experimental results showed that the risk management model has fast convergence, high predictive power, and can effectively screen defaults [6]. Scholars such as Xiong [7] aimed to build an enterprise financial shared service platform based on big data, and proposed the impact of information technology on financial shared platform services in the era of big data by comparing the traditional financial management model with the shared service model [7]. The study showed that more than 80% of enterprise accounting records were automatically generated by business-driven systems. Based on robotic process automation, Yu and Guo [8] optimized and improved the cost management process in terms of cross-system data collection, "cloud procurement platform" construction and comprehensive multi-dimensional cost analysis, which is expected to provide a reference for robotic process automation in financial shared service centre automation applications in financial shared service centre [8]. Mashrur *et al.,* [9] proposed a classification method for financial risk management tasks by considering that financial risk management can avoid losses to maximize profits, and the results showed that the method performed well in financial risk management by studying the rapidly growing financial risk management machine learning [9]. Yang *et al.,* [10] aimed to study the relationship between high frequency data and financial market impact, and used volatility as an important indicator of market risk to predict the short-term volatility of high frequency data by modeling the jump volatility of high frequency data. The results showed that the model helps to make short-term forecasts of high frequency data in financial markets [10]. Nunkoo *et al.,* [11] aimed to identify suitable autoregressive conditional duration models to capture the dynamic process of market capitalization exchange traded funds by estimating 36 ACD models for eight mid-cap ETFs with six functional forms and six error distributions [11]. The results showed that a higher degree of flexibility did not necessarily improve the goodness of fit and that the accuracy of forecasts did not always depend on the appropriateness of the model.

Many scholars have also made countless achievements in cluster analysis research. Cai *et al.,* [12] proposed a new DPC & PSO based clustering algorithm, which aims to solve the problem of the effect of parameter selection on computational density and clustering results. Experimental results showed that the method can effectively solve the problem of cluster center selection in the DPC algorithm and avoid the subjectivity of the manual selection process [12]. Long *et al.,* [13] proposed a method that uses family trees to capture the potential structure in the data by examining the density information in local neighborhoods [13]. After extensive experimentation, the algorithm was shown to be superior to DPC and spectral clustering algorithms. Yu *et al.,* [14] proposed a three-way density peak clustering method based on evidence theory and tested it on 18 datasets using three metrics (ACC, ARI and NMI) [14]. The experimental results showed that the method clustered well. Qi *et al.,* [15] used K-means clustering algorithm to study the relationship between the development of environmental industries and economic development, and used graphical analysis to analyze the correlation of electricity consumption structure and the relationship between GDP and electricity consumption of different industries [15]. The results showed that the total output of each industrial structure was proportional to the electricity consumption. Ye *et al.,* [16] proposed an improved clustering analysis algorithm based on the construction of shape features and peak features of UAV frequency hopping signal waveforms and compared it with K-means, K-means++, DBSCAN, multi-hop and autocorrelation methods [16]. Brunet-Saumard *et al.,* [17] proposed a method called “bootstrap median mean” and designed a clustering algorithm called K-bMOM considering the estimation of the mean of random variables. clustering algorithm called K-bMOM [17]. The results showed that this algorithm obtained good color quantization performance on simulated and real data sets.

In summary, scholars at home and abroad mainly analyze the financial management analysis for financial market high-frequency data or public finance, and mainly use DPC algorithm or K-means algorithm for clustering analysis. Therefore, it can be seen that the financial analysis of e-commerce companies is less often used as a research object, and no attempt has been made to improve the DPC algorithm, so this study aims to use the improved DPC algorithm to do the clustering analysis of the financial data of e-commerce companies.

3. E-commerce Financial Management Analysis Model Based on KNN Improved T-DPC Algorithm

3.1 Dimension Reduction Optimization Algorithm Model Based on T-SNE

In recent years, the rapid development of network technology has brought about the massive growth of various kinds of data, and the data shows rich and diversified development, so the real life is reflected through these high-dimensional data. The DPC algorithm is incapable of managing high-dimensional data, leading to the increasing popularity of the t-SNE algorithm in the arena of high-dimensional data presentation over the recent years. Utilizing this technology enables proficient acquisition of two-dimensional or three-dimensional coordinate points through downscaling for the efficient handling of high-dimensional data. This attribute enables users to analyze data using accessible planes or spaces. As a matter of fact, t-SNE algorithm is an improved algorithm of SNE algorithm [18]. t-SNE algorithm idea is to construct probability distribution in low-dimensional space by t-distribution with heavy tail of degree of freedom 1, as shown in Figure 1.

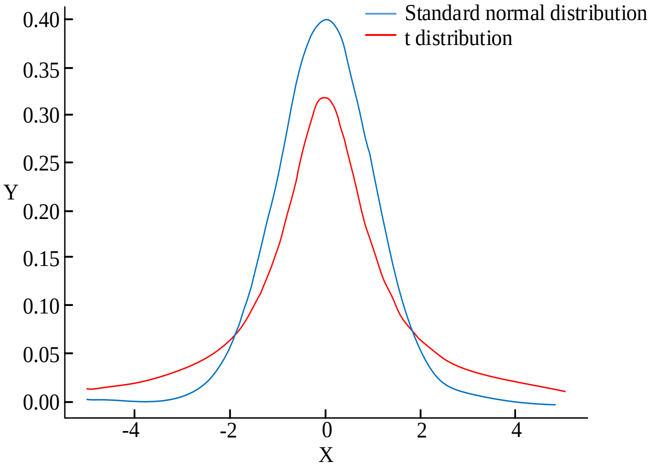


Fig. 1. T distribution and standard normal distribution function curve

From Figure 1, the SNE algorithm aims at nonlinear dimensionality reduction of high-dimensional data, while preserving more local features of the data. The core of the algorithm is to combine data with probability distribution by means of affine transformation [19]. Suppose a high-dimensional data set is $X(x_1, x_2, \dots, x_n)$, SNE can represent the high-dimensional Euclidean distance between data points by using conditional probability instead, so as to express the similarity between each object. The mathematical expression is given in Eq. (1).

$$p_{j|i} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i{}^2)}{\sum_{k\neq i}\exp(-\left\|x_i - x_k\right\|^2 / 2\sigma_i{}^2)} \tag{1}$$

In Eq. (1), x_i and x_j denote the data points, the parameter σ_i denotes the variance of the Gaussian function, i.e., the variance of x_i as the center point, whose value varies with X, and $p_{j|i}$ denotes the similarity of the data point x_i to the data point x_j, i.e., the probability that x_j becomes a sample adjacent to x_i when x_i is the Gaussian center. The value of $p_{j|i}$ is larger if the two data points are closer, and $p_{j|i}$ approaches infinity if the two data points are farther apart [20]. Suppose instead of the N n dimensional data $X(x_1, x_2, \dots, x_n)$ the data set is N d dimensional vector $Y(y_1, y_2, \dots, y_n)$, which requires d to be much smaller than n, then y_i is the subspace data point corresponding to x_i. The conditional probability of dimensionality reduction of high-dimensional data points corresponding to low-dimensional data points y_i is calculated by Eq. (2).

$$q_{j|i} = \frac{\exp(-\left\|y_i - y_j\right\|^2)}{\sum_{k\neq i}\exp(-\left\|x_i - x_k\right\|^2)} \tag{2}$$

In Eq. (2), $q_{j|i}$ represents the similarity between the data points y_i and y_j. It should be noted that the Gaussian distributed variables located in the low-dimensional space also follow the iterations of the low-dimensional variables, where the ideal case is that $q_{j|i}$ is exactly equal to $p_{j|i}$. The goal of SNE is to find a low-dimensional data space that minimizes the difference between the P_i of the original data point x_i and the Q_i of the data point y_i. The mathematical expression of Eq. (3) is then obtained by the objective function corresponding to the Kullback-Leibler scatter to obtain the result after data dimensionality reduction.

$$E_{SNE} = C = \sum_i KL(P_i \,||\, Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{3}$$

In Eq. (3), P_i represents the conditional probability distribution of the points near the high-dimensional spatial data point x_i, and Q_i represents the conditional probability distribution of the low-dimensional spatial data point y_i. Due to the asymmetry of the SNE algorithm, a large amount of gradient calculation is needed. Therefore, optimization of Eq. (3) becomes necessary. Following this optimization, the t-SNE algorithm yields the Kullback-Leibler scatter calculation formula as shown in Eq. 4).

$$\begin{cases} C = KL(P \,||\, Q) = \sum\limits_i \sum\limits_j p_{i,j} \log \dfrac{p_{ij}}{q_{ij}} \\ p_{ij} = \dfrac{p_{i|j} + p_{j|i}}{2} \\ q_{ij} = \dfrac{(1+\left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq i}(1+\left\|y_k - y_i\right\|^2)^{-1}} \end{cases} \tag{4}$$

Finally, the calculation of the gradient is completed, and the cost function can be minimized by employing the method of reduced gradient. The specific procedure involves using the initial data as a reference to produce a Gaussian distribution from a minor variable in order to acquire low-dimensional spatial data. The t-SNE algorithm is enhanced to derive the gradient calculation formula, as shown in Eq. (5).

$$\begin{cases} \dfrac{\partial C}{\partial Y} = 4\sum\limits_j (p_{ij} - q_{ij})(y_i - y_j)(1+\left\|y_i - y_j\right\|^2)^{-1} \\ Y^{(t)} = Y^{(t-1)} + \eta \dfrac{\partial C}{\partial Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)}) \end{cases} \tag{5}$$

In Eq. (5), $Y^{(t)}$ represents the solution after t iterations, η represents the learning rate, and $\alpha(t)$ represents the momentum after iterations. The introduction of this parameter can avoid the t-SNE algorithm from falling into local optimum in the optimization. Finally, the step-by-step diagram of the t-SNE algorithm shown in Figure 2 can be used to visualize the whole process of the sequence of this optimization algorithm.

From Figure 2, it is known that the t-SNE algorithm first calculates the Gaussian kernel function $d(x_i, x_j)$ for the data points x_i and x_j located in the high-dimensional space. The kernel function computes high-dimensional similarity, followed by random sampling of data points from the low-dimensional space. Conditional probabilities within the low-dimensional space are calculated, and then matched with those within the high and low dimensions to compare similarity between the two. Finally, the Kullback-Leibler scattering is utilised for the optimization search calculation aided by the gradient descent method.

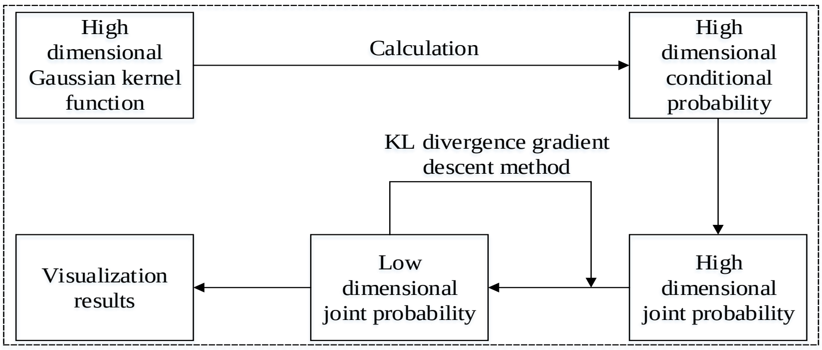

High dimensional Gaussian kernel function
Calculation
High dimensional conditional probability
KL divergence gradient descent method
Visualization results
Low dimensional joint probability
High dimensional joint probability

Fig. 2. t-SNE algorithm step diagram

3.2 Design of DPC Algorithm Improved Based on K-neighborhood Idea

The DPC algorithm aims to extract the class cluster centers, i.e., the class cluster centers are considered as data center points with high density, and the density of the class cluster centers is considered to be higher than the density of their neighboring data points, and also the class cluster centers are considered to be further away from other denser data points than other less dense data points. Therefore, in clustering analysis, it is necessary to find the centroid of the data set and then categorize the remaining data points. The DPC algorithm calculates the data point distance using Eq. (6).

$$d_{ij} = \sqrt{\sum_{k=1}^{N} (X_i(k) - X_j(k))^2} \tag{6}$$

In Eq. (6), d_{ij} represents the distance between the data point X_i and the data point X_j, and $X_i(k)$ represents the data value of the data point X_i in the Kth dimension. After that, the truncation distance d_c is calculated by sorting all data points d_{ij} in ascending order and then artificially taking a suitable percentage from 0.5% to 5%, and the truncation distance is the ascending order of d_{ij} multiplied by that percentage. The local density ρ_i is the main parameter of the DPC algorithm, which is defined as the number of data points X_i whose Euclidean distance is less than the truncation distance. If the size of the data set is large, the clustering effect is less affected by the truncation distance and its mathematical expression Eq. (7).

$$\rho_i = \sum_{i \neq j} \chi(d_{ij} - d_c) \tag{7}$$

In Eq. (7), χ is a constant coefficient, which takes the value of 0 or 1. However, if the size of the data set is small, the clustering effect will be affected by the truncation distance, which is mathematically expressed in Eq. (8).

$$\rho_i = \sum_{i \neq j} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{8}$$

Define δ_i to denote the minimum distance among the data nodes with a local density greater than X_i, and its expression is Eq. (9). Thus, the local density of each data point and its minimum distance are obtained.

$$\delta_i = \min_{j \cdot \rho_{j>} \rho_i} (d_{ij}) \tag{9}$$

The decision diagram can be constructed from the local density and the minimum distance of the data points. Figure 3(a) displays the data distribution, which indicates that the data is centered around point 1 and point 10. Similarly, Figure 3(b) shows the decision graph, revealing that both point 1 and point 10 are subject to large data of ρ_i and δ_i respectively.

The remaining data points are then assigned and the class cluster boundaries and outliers are identified. The method of assigning the remaining data points means that each data point to be assigned is assigned to the center of its nearest and locally densest cluster. Cluster boundaries are determined by comparing the distance between the assigned data point and the cluster center with the truncation distance, and if the former is less than the truncation distance d_c , then the point is considered a boundary point. The outlier is the point that has a greater density than the point with the maximum density among the cluster boundary points. It can be seen that the DPC algorithm has the advantage of simple and efficient application, but its disadvantage is also obvious, that is, the truncation distance in the DPC algorithm is set manually, and this setting method will affect the density calculation and the final clustering effect. The K-nearest neighbor algorithm is one of the common algorithms for classification tasks, and it is often used for object evaluation and prediction [21]. The formula for calculating the local density after improvement by KNN-DPC algorithm is Eq. (10).

$$\rho_i = \sum_{j \in KNN(i)} e^{-d_{ij}} \tag{10}$$

In Eq. (10), $KNN(i)$ represents the set consisting of k nearest neighbor data points of data poin t X_i, and d_{ij} represents the Euclidean distance between data points X_i and X_j. It can be seen from Eq. (10) that the closer the data point is to the X_i k nearest neighbor data points, the greater its local density, so that the original density calculation can be narrowed down by this improved formula, i.e., from the whole data set range to the data point next to the k nearest neighbor, so as to reflect the local information of the data point X_i more accurately. The KNN-DPC algorithm for outlier points is defined by Eq. (11).

$$\begin{cases} k_{dist}(i) = \max_{j \in KNN(i)} \{d_{ij}\} \\ threshold = \dfrac{1}{N} \sum_{i=1}^{N} k_{dist}(i) \\ Outlier = \{o \mid k_{dist}(o) > threshold\} \end{cases} \tag{11}$$

In Eq. (11), k_{dist} represents the KNN distance of each sample and $threshold$ represents the threshold value. If the KNN distance is greater than the threshold, then the data point is an outlier. Then the data is assigned by the outlier, and if there is still an unassigned data point after the assignment is completed, then the point is a noise point [22].

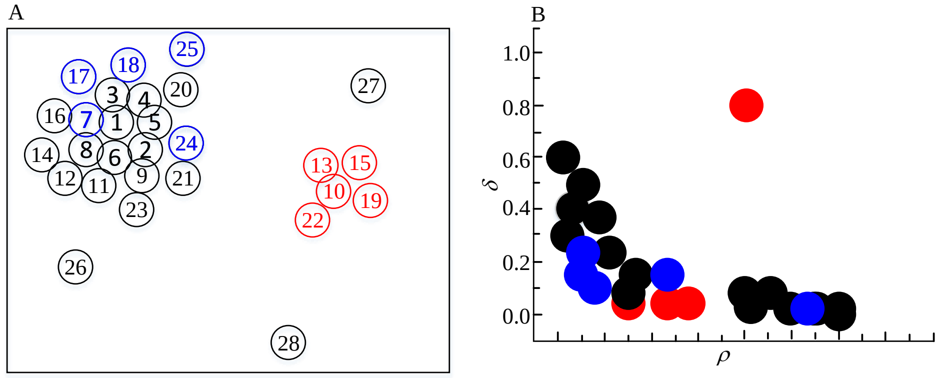


Fig. 3. Example of decision graph constructed from local density and minimum distance

4. Analysis of the Clustering Results of E-commerce Financial Data Based on KNN-T-DPC Algorithm

4.1 KNN-T-DPC Algorithm Performance Comparison Utility Analysis

Three artificial datasets, Aggregation, D31 and R15, and five UCI datasets, PID, Wine, Iris, Waveform and Seed, were selected as performance test samples of the T-DPC algorithm to verify its effectiveness and superiority. Meanwhile, the F-measure is used to evaluate the clustering effect of T-DPC algorithm, and the efficiency of the algorithm is evaluated by comparing the running time of T-DPC algorithm and DPC algorithm. The clustering effects of the two algorithms are shown in Figure 4.

As can be seen from Figure 4, the performance of T-DPC and DPC algorithms on F-measure metrics is not very different, and the difference is kept within 0.02, with a minimum of 0. The reason is that the size and complexity of these three artificial datasets are not high, and they do not need to do data dimensionality reduction by t-SNE, so there is basically no difference in accuracy between the two algorithms. However, for the last five high dimensional UCI datasets, they need to do dimensionality reduction by t-SNE in order to be more suitable for DPC to do clustering analysis. From Figure 4, the data illustrates that the suggested model has exhibited a 16.7% and 3.07% enhancement as opposed to the unoptimized DPC model in the PID dataset and Wine dataset, respectively, regarding the F metric. For the algorithm run time calculation, the average of the 30 algorithm run times of the selected stocks was used as the final run comparison time in order to improve the experimental accuracy. Table 1 shows the results of computing time comparison between the two algorithms.

As can be seen from Table 1, the running times of the T-DPC algorithm on the artificial data sets Aggregation, D31 and R15 are 16.2s, 193.1s and 10.3s respectively, which are reduced by 5s, 115.1s and 2.9s respectively compared to the DPC algorithm. On the standard Waveform and PID datasets, the time is reduced by 331.6s and 1.3s respectively, while there is no change on the standard Wine and Iris datasets, as these two datasets are smaller in size compared to the other three standard datasets and therefore cannot reflect the advantages of the T-DPC algorithm's data dimensionality reduction. Therefore, it can be seen from Table 2 that the T-DPC algorithm is more suitable for large datasets compared to small datasets, and its algorithm runs more efficiently, making it more suitable for analyzing the large and complex data generated by e-commerce companies.

Considering the influence of the artificially set truncation distance of the DPC algorithm, the study used the K-nearest neighbor algorithm to improve the DPC algorithm to obtain the KNN-DPC algorithm. Three artificial datasets, Aggregation, D31 and R15, and three UCI datasets, Iris, Parkinson and Seed, were selected as test samples, and the DPC algorithm, DBSCAN algorithm, AP algorithm and K-means algorithm were used as comparison algorithms, and three metrics, Acc, AMI and ARI, were used as algorithms. Table 1 shows the results of the metric data obtained by the five clustering algorithms with the artificial datasets as test samples.

As can be seen in Table 2, the three metrics Acc, AMI and ARI obtained from the clustering analysis of the KNN-DPC algorithm have the highest values in the three artificial datasets tested, indicating that the algorithm outperforms the other four algorithms. For the performance difference between KNN-DPC and DPC algorithms, on the Aggregation dataset, the KNN-DPC algorithm has values of 0.998, 0.994 and 0.997 on the three metrics, which are 0.001, 0.002 and 0.001 better than the DPC algorithm, respectively; on the D31 dataset, the KNN-DPC algorithm has three evaluation metrics values of 0. 963, 0.943 and 0.997, which are 0.006, 0.007 and 0.006 better than the DPC algorithm respectively; for the R15 dataset, the KNN-DPC algorithm has values of 0.997, 0.994 and 0.993 for the three metrics, which are 0.005, 0.006 and 0.011 better than the DPC algorithm respectively. Table 3 shows the five clustering algorithms using the UCI dataset as a test sample to obtain the results of the metric data.

As can be seen from Table 3, the KNN-DPC algorithm did not obtain optimal values only for the AMI metric on the Parkinson dataset, while it was optimal in all other cases. For example, on the Iris dataset, the KNN-DPC algorithm achieves 0.967, 0.907 and 0.914 for the Acc, AMI and ARI metrics respectively, which are higher than those of the DPC, AP, DBSCAN and K-means algorithms, with an improvement of 0.08, 0.14 and 0.194 respectively over the DPC algorithm; and on the seed Table 4 shows the running time of the five clustering algorithms when tested on the UCI dataset.

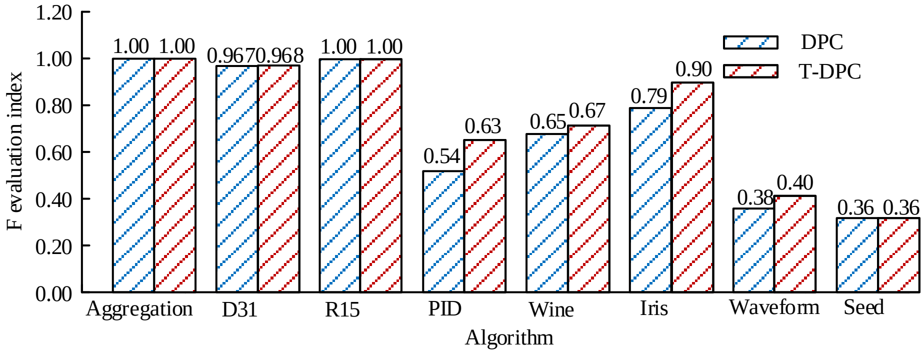


Fig. 4. Clustering effect of two algorithms on F-measure index

Table 1
Comparison results of operation time of two algorithms

Dataset	DPC	T-DPC
Aggregation	21.2s	16.2s
D31	308.2s	193.1s
R15	13.2s	10.3s
PID	22.5s	21.2s
Wine	4.2s	4.2s
Iris	3.6s	3.6s
Waveform	843.2s	511.6s
Seed	3.5s	3.3s

Table 2
Index data results of five clustering algorithms on artificial data sets

Algorithm	Aggregation			D31			R15		
	Acc	AMI	ARI	Acc	AMI	ARI	Acc	AMI	ARI
KNN-DPC	0.998	0.994	0.997	0.975	0.963	0.943	0.997	0.994	0.993
DPC	0.997	0.992	0.996	0.969	0.956	0.937	0.992	0.987	0.982
AP	0.760	0.795	0.702	0.956	0.948	0.986	0.981	0.975	0.978
DBSCAN	0.986	0.986	0.988	0.954	0.926	0.928	0.987	0.964	0.979
K-means	0.765	0.795	0.687	0.683	0.769	0.568	0.995	0.993	0.992

Table 3
Index data results of five clustering algorithms on UCI dataset

Algorithm	Iris			Parkinson			Seed		
	Acc	AMI	ARI	Acc	AMI	ARI	Acc	AMI	ARI
KNN-DPC	0.967	0.907	0.914	0.810	0.171	0.257	0.921	0.759	0.802
DPC	0.887	0.767	0.72	0.610	0.201	0.027	0.900	0.717	0.734
AP	0.899	0.754	0.762	0.654	0.156	0.154	865	0.675	0.675
DBSCAN	0.856	0.758	0.765	0.685	0.263	0.245	0.846	0.678	0.634
K-means	0.854	0.635	0.654	0.652	0.254	0.039	0.867	0.657	0.763

Table 4
Running time of five clustering algorithms on UCI dataset

Dataset	KNN-DPC	DPC	AP	DBSCAN	K-means
Iris	0.032	0.049	0.534	0.063	0.012
Parkinson	0.092	0.172	1.236	0.076	0.015
Seed	0.152	0.185	0.265	0.046	0.016
WDBC	0.119	0.292	1.254	0.086	0.015
Wine	0.142	0.248	0.835	0.093	0.012
Ionosphere	0.131	0.164	0.542	0.053	0.026
Segmentation	0.655	0.806	0.735	0.125	0.059

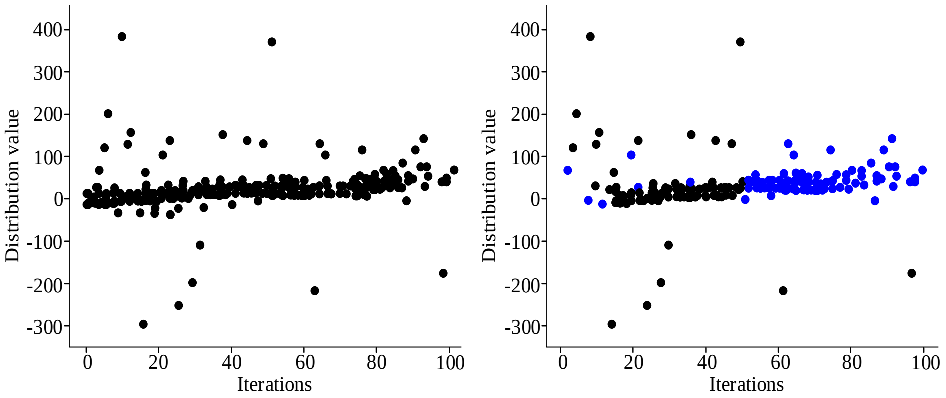
As shown in Table 4, the running time of the KNN algorithm on the seven UCI datasets is 0.032 s, 0.092 s, 0.152 s, 0.119 s, 0.142 s, 0.231 s and 0.655 s, which is 0.017 s, 0.080 s, 0.030 s, 0.173 s, 0.146 s and 0.655 s faster than the DPC algorithm in terms of running time. 0.033 s and 0.151 s. Therefore, the data in Table 4 can show that the KNN-DPC algorithm works more efficiently than the DPC algorithm.

4.2 Analysis of the Results of E-commerce Financial Data Using KNN-T-DPC

The study utilizes financial data from 40 listed companies during the third quarter of 2021 as the analytical focus. To reflect the comprehensive profitability of listed companies, indicators such as profitability, growth ability, debt-paying ability and operational capability are selected to evaluate the financial position of enterprises. To classify and select financially sound and financially poor enterprises, the KNN-T-DPC algorithm is utilized. Figure 5 reflects the results of the clustering analysis of KNN-T-DPC algorithm.

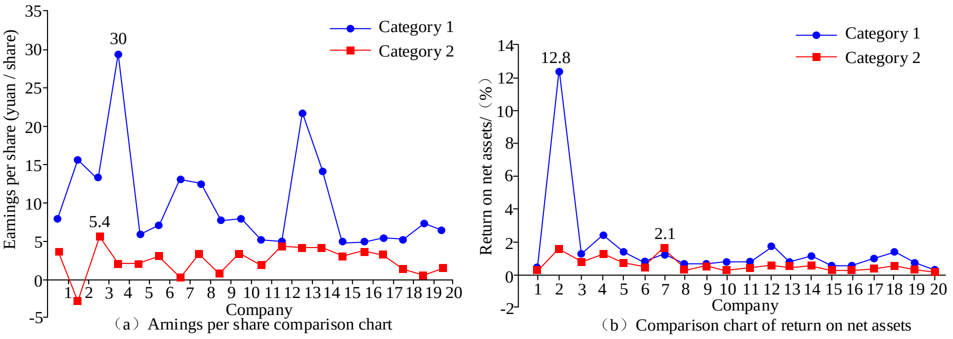
From Figure 5(a), it can be seen that only a few companies have differences between their financial situation and other companies, while most companies are clustered into the same category, showing a high similarity. The reason is that there are more listed companies in recent years, and these listed companies can operate stably under national macroeconomic policies, and more financial indicators are selected to perform cluster analysis on large-capacity data samples, so financially unhealthy enterprises are not selected, and the classification is not obvious. Figure 5(b) shows the clustering results of only 5 selected financial indicators, and the stock data of listed companies were considered in the clustering, so their financial indicators are enterprise net profit, earnings per share value, return on net assets, operating income value and net assets per share, and the final sample was clustered into 2 categories. Figure 6 shows the comparative graphs of earnings per share and return on net assets for the two categories.

From Figure 6(a), the stock returns of the first category are greater than the second category, indicating that the listed companies in the first category are more profitable in the stock market. From Figure 6(b), it can be seen that the highest return on net assets is 12.8% and the lowest is 0.5% for the first category of companies. The highest NAV of the second category enterprises is 2.1%, and most of the first category enterprises have NAVs greater than or equal to the second category. Therefore, in summary, it can be seen that the first category of enterprises has strong overall strength, while the second category of enterprises has lower figures in both indicators, and some enterprises have negative values in earnings per share, which indicates that these enterprises have poorer performance and weaker profitability compared with the first category of enterprises, and the company’s development ability and prospects are less optimistic.



(a)Select 10 characteristic index clustering results (b)Select 5 characteristic index clustering results

Fig. 5. Clustering analysis results of KNN-T-DPC algorithm



（a）Arnings per share comparison chart （b）Comparison chart of return on net assets

Fig. 6. Comparison chart of earnings per share and return on net assets

5. Conclusion

E-commerce enterprises fall under the emerging tertiary industry, which is intensely competitive. Therefore, the examination of their financial management is of great significance. The initial step of this investigation involved the implementation of the t-SNE algorithm to reduce the dimensionality of abundant complex high-dimensional data present in e-commerce companies. The K-nearest neighbor algorithm was subsequently introduced to enhance the DPC algorithm, culminating in the KNN-t-DPC algorithm. Ultimately, the superior algorithm will undergo further analysis of its performance and practical application. The t-SNE optimized DPC algorithm proposed by the research institute has higher F-measure metrics than the standard DPC algorithm in multiple datasets (PID, Wine, Iris, Waveform, etc.). When analyzing the application of financial data in e-commerce enterprises, it was found that the KNN-T-DPC algorithm lacks significant classification performance when clustering more indicators. Despite achieving a high level of prediction accuracy, the T-DPC model proposed by the research institute has a large number of cluster centers due to its high data dimension. In future research, it is possible to carry out pertinent pruning once the cluster centers have been obtained.

Author Contributions

Conceptualization, Y.W. and Y.S.; methodology, Y.W. and Y.S.; software, Y.W. and Y.S.; validation, Y.W. and Y.S.; formal analysis, Y.W. and Y.S.; investigation, Y.W. and Y.S.; resources, Y.W. and Y.S.; data curation, Y.W. and Y.S.; writing—original draft preparation, Y.W. and Y.S.; writing—review and editing, Y.W. and Y.S.; visualization, Y.W. and Y.S.; supervision, Y.W. and Y.S.; project administration, Y.W. and Y.S.; funding acquisition, Y.W. and Y.S. All authors have read and agreed to

the published version of the manuscript. Authorship must be limited to those who have contributed substantially to the work reported.

Funding

This research received no external funding .

Data Availability Statement

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research is supported by: the Science and Technology Support Programs of Langfang City (2021011063) and the Higher Education Scientific Research Planning Projects of the North China Institute of Science and Technology (HKJYZD202004).

References

[1] Mai, K., Saeed, W., & Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: logistic regression, decision tree, and random forest. Applied Soft Computing, 118(1), 108491. https://doi.org/10.1016/j.asoc.2022.108491

[2] Munshi, M., Shrimali, T., & Gaur, S. (2022). A review of enhancing online learning using graph-based data mining techniques. Soft Computing, 26(12):5539-5552. https://doi.org/10.1007/s00500-022-07034-7

[3] Safarnejad, L., Xu, Q., Ge, Y., & Chen, S. (2021). A multiple feature category data mining and machine learning approach to characterize and detect health misinformation on social media. IEEE Internet Computing, 25(5), 43-51. https://doi.org/10.1109/MIC.2021.3063257

[4] Pistikou, V., Tsanana, E., & Poufinas, T. (2021). A Financial Analysis Approach on the Impact of Economic Interdependence on Interstate Conflicts. Theoretical Economics Letters, 11(5),947-961. https://doi.org/10.4236/tel.2021.115060

[5] Cohen, S., Manes-Rossi, F., Brusca, I., & Caperchione, E. (2021). Guest editorialHappy endings and successful stories in public sector financial management: a lesson drawing perspective. International Journal of Public Sector Management, 34(4), 393-406. https://doi.org/10.1108/IJPSM-05-2021-347

[6] Qiu, W. (2021). Enterprise financial risk management platform based on 5G mobile communication and embedded system. Microprocessors and Microsystems, 80, 103594. https://doi.org/10.1016/j.micpro.2020.103594

[7] Xiong, L. (2021, April). Enterprise financial shared service platform based on big Data. In *Journal of Physics: Conference Series* (Vol. 1852, No. 3, p. 032004). IOP Publishing. https://doi.org/10.1088/1742-6596/1852/3/032004

[8] Yu, L. Q., & Guo, F. X. (2020). Research on Cost Management Optimization of Financial Sharing Center Based on RPA. Procedia Computer Science, 166, 115-119. https://doi.org/10.1016/j.procs.2020.02.031

[9] Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. *IEEE Access*, *8*, 203203-203223. https://doi.org/10.1109/ACCESS.2020.3036322

[10] Yang, R., Yu, L., Zhao, Y., Yu, H., Xu, G., Wu, Y., & Liu, Z. (2020). Big data analytics for financial Market volatility forecast based on support vector machine. *International Journal of Information Management*, *50*, 452-462. https://doi.org/10.1016/j.ijinfomgt.2019.05.027

[11] Nunkoo, H. B. S., Gonpot, P. N., Sookia, N. U. H., & Ramanathan, T. V. (2022). Autoregressive conditional duration models for high frequency financial data: an empirical study on mid cap exchange traded funds. *Studies in Economics and Finance*, *39*(1), 150-173. https://doi.org/10.1108/SEF-04-2021-0146

[12] Cai, J., Wei, H., Yang, H., & Zhao, X. (2020). A novel clustering algorithm based on DPC and PSO. *IEEE Access*, *8*, 88200-88214. https://doi.org/10.1109/ACCESS.2020.2992903

[13] Long, Z., Gao, Y., Meng, H., Yao, Y., & Li, T. (2022). Clustering based on local density peaks and graph cut. Information Sciences, 600, 263-286. https://doi.org/10.1016/j.ins.2022.03.091

[14] Yu, H., Chen, L., & Yao, J. (2021). A three-way density peak clustering method based on evidence theory. Knowledge-Based Systems, 211, 106532. https://doi.org/10.1016/j.knosys.2020.106532

[15] Qi, Y., Ren, J., Sun, N., & Yu, Y. (2021, July). Application of clustering algorithm by data mining in the analysis of smart grid from the perspective of electric power. In Journal of Physics: Conference Series (Vol. 1982, No. 1, p. 012018). IOP Publishing. https://doi.org/10.1088/1742-6596/1982/1/012018

[16] Ye, J., Zou, J., Gao, J., Zhang, G., Kong, M., Pei, Z., & Cui, K. (2021). A new frequency hopping signal detection of civil UAV based on improved k-means clustering algorithm. IEEE Access, 9, 53190-53204. https://doi.org/10.1109/ACCESS.2021.3070491

[17] Brunet-Saumard, C., Genetay, E., & Saumard, A. (2022). K-bMOM: A robust Lloyd-type clustering algorithm based on bootstrap median-of-means. Computational Statistics & Data Analysis, 167, 107370. https://doi.org/10.1016/j.csda.2021.107370

[18] Kobak, D., & Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nature biotechnology, 39(2), 156-157. https://doi.org/10.1038/s41587-020-00809-z

[19] Bajal, E., Katara, V., Bhatia, M., & Hooda, M. (2022). A review of clustering algorithms: comparison of DBSCAN and K-mean with oversampling and t-SNE. Recent Patents on Engineering, 16(2), 17-31. https://doi.org/10.2174/1872212115666210208222231

[20] Gove, R., Cadalzo, L., Leiby, N., Singer, J. M., & Zaitzeff, A. (2022). New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation. Visual Informatics, 6(2), 87-97. https://doi.org/10.1016/j.visinf.2022.04.003

[21] Almomany, A., Ayyad, W. R., & Jarrah, A. (2022). Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study. Journal of King Saud University-Computer and Information Sciences, 34(6), 3815-3827. https://doi.org/10.1016/j.jksuci.2022.04.006

[22] Trieu, N. M., & Thinh, N. T. (2022). A study of combining knn and ann for classifying dragon fruits automatically. Journal of Image and Graphics, 10(1), 28-35. https://doi.org/10.18178/joig.10.1.28-35