Research Article

Analysis of Internet Marketing Forecast Model Based on Parallel K-Means Algorithm

Xiaolei Chen¹ and Sikun Ge²

¹*School of Manage, Guangzhou College of Commerce, Guangzhou, Guangdong 511363, China*

²*Department of Philosophy and Culture, Guangzhou Administration Institute, Guangzhou, Guangdong 510030, China*

Correspondence should be addressed to Xiaolei Chen; 2009853gbb30006@student.must.edu.mo

Received 26 October 2021; Accepted 19 November 2021; Published 2 December 2021

Academic Editor: Miaochao Chen

Copyright © 2021 Xiaolei Chen and Sikun Ge. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Based on the parallel K-means algorithm, this article conducts in-depth research on the related issues of marketing node detection under the Internet, including designing a new Internet marketing node detector and a location summary network based on FCN (Full Convolutional Network) to input the preprocessing of the node and verify its performance under the data sets. At the same time, to solve the problem of insufficient data sets of Internet marketing nodes, the Internet data sets are artificially generated and used for detector training. First, the multiclass K-means algorithm is changed to two categories suitable for Internet marketing node detection: marketing nodes and background categories. Secondly, the weights in the K-means algorithm are mostly only applicable to target detection tasks. Therefore, when processing Internet marketing node detection tasks, the K-means algorithm is used to regress the training set and calculate 5 weights. During the simulation experiment, the weight calculation formula is used to calculate the weight of the feature term. The basic idea is that if a feature word appears more often in this document but less frequently in other nodes, the word will be assigned higher. At the same time, this article focuses on k . Some shortcomings of the mean clustering algorithm have been specifically improved. By standardizing the data participating in the clustering, the data participating in the clustering is transformed from an irregular distribution to a cluster-like distribution, thereby facilitating the clustering process. The density is introduced to determine the initial center of the cluster, and the purity metric is introduced to determine the appropriate density radius of the cluster center, to achieve the most effective reduction of the support vector machine training samples.

1. Introduction

The scale of the Internet is increasing day by day, and at the same time, a huge amount of relevant data is generated. The traditional Internet marketing forecasting technology is constrained by the computer's performance and programming model, resulting in a bottleneck, and it is significantly helpless when processing these data [1]. Internet marketing forecasts are designed to automatically search for information with special relationships hidden in a large amount of data. In the case of massive data and high-dimensional data, a single processor is limited by computing power and memory capacity, so a solution for parallel processing by multiple processors has been proposed [2–5]. The most common idea is to divide a large-scale data set into multiple

data subsets that are sufficient for single-machine processing and distribute these subsets to each single-processor node for processing. When each node assigns itself to the processing of the data subset after completion, the respective results are summarized and merged, and the final result is the processing result of the entire data set. Compared with a single processor, the parallel data processing mode of multiple computing nodes on multiple processors can significantly improve the efficiency of Internet marketing forecasting [6–8].

The main idea of Internet marketing forecasting is to distribute computing tasks on a virtual resource pool composed of many computers so that various application systems can obtain computing power, storage space, and various software services as needed. The research

assumption in this paper is to deploy a cloud computing platform and parallelize the clustering algorithm MapReduce while optimizing details such as data segmentation, task allocation, parallel processing, and fault tolerance. Due to the variety of clustering algorithms, this article only starts with the K-means clustering algorithm, combines the traditional K-means algorithm with the Canopy algorithm, and parallelizes the above two algorithms according to the MapReduce programming model [9]. The improved algorithm is applied to the Internet cloud computing platform. Through the SogouC data set and Chinese data set, the marketing node clustering processing is performed. The experiment shows that the parallelized MapReduce algorithm greatly improves the running speed. Therefore, the Canopy algorithm is more capable of clustering large-scale data sets than the K-means algorithm [10–12].

Aiming at the shortcomings of the KM algorithm and BKM algorithm in clustering, this paper proposes a cooperative dichotomous K-means algorithm (CBKM) based on the idea of cooperative clustering. The algorithm is mainly divided into three stages: overall clustering, cooperative clustering, and fusion. This algorithm is realized by the intermediate cooperation of synchronization through KM in the process of BKM generating CF tree. By introducing the concept of similar histograms, it can intuitively reflect the adhesion of elements between clusters. According to similar histograms of similar subcategories, the fusion factor of the two subcategories is calculated, the two clusters with the largest fusion factor values are merged, and the clusters are updated. In summary, the Euclidean distance is still selected as the main similarity measure of function clustering. The clustering results produced by this process can effectively avoid the generation of cluster fragments. Because the intersection of subcategories is merged and clustered, it effectively improves the K-means algorithm to be affected by the initial clustering center. The most global solution is not a local optimal solution.

2. Related Work

With the development of Internet technology, the network has an increasing impact on people's lives. The information on the network mainly exists in the form of voice, node, and so on. Among them, the amount of information with nodes as carriers is the largest. How to effectively mine marketing node information has become an important research topic in recent years. Marketing node clustering can organize and manage a large number of marketing nodes, which largely solves the problems caused by information explosion and information clutter.

Ghayekhloo et al. [13] proposed the K-means algorithm, UPGMA, CLARANS, and other algorithms that have not received attention. They have only paid attention to it in recent years and achieved greater development. In the new century, with the rapid development of the Internet, digital libraries and the widespread use of e-commerce, the scale of the marketing node set has gradually expanded. Some new algorithms have been proposed, such as FTC, HDFA, and FCA. The current research on clustering algorithms is getting deeper and more

refined. According to the characteristics of clustering, it can be divided into division-based, hierarchical-based, model-based, density-based, and grid-based methods. Pan et al. [14] proposed the SWT (Stroke Width Transformation) feature specifically applicable to marketing nodes. However, artificially defined features such as MSER and SWT are not suitable for scene nodes with extremely complex backgrounds. Since then, with the rapid advancement of deep learning technology, more and more researchers have introduced deep learning into the task of Internet marketing node detection. Internet marketing node detection includes two steps: marketing node area positioning and node identification. A 3×3 size horizontal difference filter is used to obtain the vertical edge of the scene node. Then, the separated node parts are connected with the low-frequency enhanced spatial domain filtering technology. Finally, some a priori heuristic knowledge, such as area ratio and aspect ratio, filter out nonmarketing node areas. Mao et al. [15] proposed a method for extracting marketing nodes in videos, using multiresolution node input to solve the problem of node sizes in nodes and then using an improved Sobel operator to obtain node edges and perform node processing. The method based on edge detection has a relatively accurate detection effect for nodes with a relatively standard marketing node format, such as printing fonts in nodes.

Veerasha et al. [16] improved the system to improve classification accuracy. They proposed to use top-down, layer-by-layer classification. According to the concept dictionary, the synonym concept was mapped to a single-concept word, and the dimension of the feature vector was reduced to reduce the amount of calculation. It is based on the SWT features proposed by people that are specifically applicable to marketing nodes, as well as the improvement and expansion of the above features. Most of the nodes in the same row have the same brightness and color. Based on this feature, Huang et al. [17] applied the MSER feature to scene marketing node positioning. The MSER feature extraction algorithm uses the idea of the watershed to extract the areas with roughly the same brightness value in the nodes as node features. However, the MSER feature extraction algorithm has very limited effects in complex situations such as uneven illumination. Researchers propose SWT (Stroke Width Transformation) features that are specifically applicable to marketing nodes. The SWT feature extraction algorithm considers a node to have a basic feature; that is, each stroke in the node has roughly the same stroke width. SWT uses the first-order difference method to calculate the stroke width value in the node and combines pixels with roughly the same stroke width value into a connected domain to find the node position [18–20].

3. Internet Marketing Prediction Model Construction Based on Parallel K-Means Algorithm

3.1. Parallel K-Means Algorithm Flow. The K-means algorithm's (KM) clustering effect is affected by the initial cluster center, it is difficult to have a unified standard for the

selection of k value, the selection of the initial cluster center will have a greater impact on clustering, and the existence of outliers will cause many problems. It is difficult to find the global optimal solution. Clustering is a flexible automatic processing mechanism that does not need to manually mark node categories in advance and does not need to be trained. It is an unsupervised learning method [21–23]. The formal description is as follows: given a data set X and the distance measure or similarity d between the data points, find the division from the data set X . Figure 1 is the process framework of the parallel K-means algorithm.

We perform node preprocessing on the target marketing node set. This stage includes operations such as word segmentation, removal of staying words, and stemming. Next, we extract the feature items, reduce the dimension of the feature items, select the marketing node representation method, construct the feature space, form the document vector, select the distance and similarity measure, and finally perform the marketing node clustering process to obtain the clustering result:

$$\left(\frac{z+\lambda}{a_1}\right)^2 - \left(\frac{x+y}{a_2}\right)^2 = 1, \quad (1)$$

$$\begin{bmatrix} x_{11} & x_{12} & x_{1n} \\ x_{21} & x_{22} & x_{2n} \\ x_{m1} & x_{m2} & x_{mn} \end{bmatrix} = x(i, j).$$

The hierarchical methods decompose the data set hierarchically through splitting or cohesion. First, we put all the data points in the data set into one cluster and then split this cluster into two clusters according to a certain principle:

$$r(i, j) = \frac{(x(i, j) - \min(x))}{(\max(x) - x(i, j))}, \quad i, j = 1, 2, \dots, m, \quad (2)$$

$$r(s, t) = \frac{(\max(x) - x(s, t))}{(x(s, t) - \min(x))}, \quad s, t = 1, 2, \dots, n.$$

Generally, the principle of optimal objective function value is adopted and then after multiple iterations of splitting until the convergence function is stable. This top-down hierarchy-based method is called the split method. Then, the cohesion method first separates each data point into categories and then merges the categories one by one according to a certain principle until a certain termination condition is met:

$$x(\text{mean}) = \lim_{m \rightarrow \infty} \sum_{i,j=1}^m \frac{x(i, j) - 1}{r(i, j) - 1},$$

$$w(i, j) = \frac{(x(i, j) - \bar{x})(x(i, j) + \bar{x})}{\sum x(i) \times x(j)} \times 100\%, \quad i, j = 1, 2, \dots, m. \quad (3)$$

Marketing node mining is not a simple task. It deals with unstructured or semistructured marketing node data, most

of which are vague and lack a definite form. Therefore, to complete the process of text mining, it is necessary to combine multiple disciplines, including marketing node analysis. Therefore, the traditional Internet marketing forecasting technology cannot be fully applied to the field of marketing node mining. In order to solve this problem, the marketing node data set needs to be fully and effectively preprocessed:

$$c = \Pr(c|o) \times \Pr(o) \times \text{IoU} = \Pr(c) \times \text{IoU},$$

$$l = \alpha \sum_{i=0}^{S^2} \sum_{j=0}^B \left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + (x_j - \bar{x})^2 + (y_j - \bar{y})^2 \right]. \quad (4)$$

The basic idea is to estimate the probability of correlation between marketing nodes and queries and rank all marketing nodes according to the correlation probability. In the case of a given query Q , P represents the probability that the marketing node is related to the query, and $P-R$ represents the probability that the marketing node is not related to the query:

$$\min \sum_{i=1}^N a_i = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i y_j a_j y_i \times k(x_i^2, x_j^2). \quad (5)$$

The purpose of the clustering algorithm is to cluster similar data into one category. Therefore, to measure the degree of similarity, the definition of similarity, also called proximity, is introduced to describe it. Contrary to the distance measure, the smaller the value of the similarity measure is, the smaller the similarity between individuals and the greater the difference is.

$$\begin{cases} V_{(i,f)} = \frac{J_i^f + J_i^{f-1}}{T}, \\ V_i = (V_{(i,i)}, \dots, V_{(i,f)}). \end{cases} \quad (6)$$

Cosine similarity calculates the cosine value of the angle between two vectors in the vector space. The effective value of the cosine value in the vector space ranges from 0 to 1. 0 means the two vectors are completely independent, and 1 means that the directions of the two vectors are exactly the same. Moreover, these shortcomings can be overcome by standardizing the function in advance or clustering the specific derivative of the function. Compared with the distance measurement, cosine similarity does not focus on distance and length but the difference in direction.

3.2. Internet Data Clustering. Internet marketing forecasting refers to extracting knowledge that is valuable to people from a large amount of random, incomplete, fuzzy, and noisy data, making full use of nodes and signal processing, statistics, pattern recognition, machine learning, theories, and methods in various fields such as information retrieval and database technology:

FIGURE 1: Parallel K-means algorithm process framework.

$$L_{(emg,k)} = \sum_{f(emg,k) \in A_{(emg,k)}} (f_{(emg,n)} + C_{(emg,k)}), g(m, t) = \frac{n \times [h(m, t)/k(i, t)]}{\lim_{n \rightarrow \infty} \sum_{i=1}^n (h(i, t)/k(i, t))}. \quad (7)$$

Marketing node mining is an important research content in Internet marketing forecasting, which mainly includes marketing node classification and marketing node clustering. Clustering is to divide a data set into different classes or clusters according to a certain standard (usually a distance criterion) so that the similarity of data objects in the same cluster is as large as possible. At the same time, the differences of data objects that are not in the same cluster are also as large as possible. In other words, after clustering, the data of the same category should be gathered together as much as possible, while different data should be separated as much as possible. Cluster analysis is a basic method for data analysis, and it has been widely studied and applied in many fields, such as market or customer segmentation. Figure 2 shows the classification of Internet marketing prediction nodes.

In the test classification stage, the test sample set S and the category of the test sample are used to evaluate the degree of consistency between the classifier and the objective function. Each element ink is also marked with its corresponding category $y \in C$. The specific process is to classify each test sample with the previously obtained classifier and finally evaluate the classifier according to the classification result: if the classification performance on the test set fails to reach the predetermined goal, then return to the training phase and use more samples to relearn or modify the learning algorithm; if the test classification performance reaches the predetermined goal, then the classifier can be applied to the new marketing node to be classified for

automatic classification to suppose that there is a group of M particles in a D -dimensional target search space. At the same time, this method also has shortcomings and deficiencies. The clustering quality is greatly affected by the lowest level of granularity, and the clustering accuracy is not very impressive. In the group, each particle corresponds to a solution of the optimization problem, a fitness value is determined by the objective function, and the speed is used to determine the motion of the particle in the solution space. In each iteration, the particle will track the optimal solution currently found by itself and the population and iteratively search until the optimal solution is finally obtained.

3.3. Marketing Forecast Classification Metrics. In the process of marketing forecast classification, when faced with high-dimensional data and large-scale data, the traditional stand-alone computing model cannot provide efficient processing results per unit time. Due to the bottleneck in the computing speed of the single-processor CPU, the technology of parallel processing by multiple processors has been widely used. Traditional parallel computing is to perform operations on high-performance computing systems such as parallel computers or distributed computers. Common parallel computing includes a cluster system composed of multiple single processors. This system relies on improving the performance of a single processor to improve the overall performance of the parallel system. As well as network performance, this parallelism is instruction-level parallelism.

FIGURE 2: Classification of Internet marketing forecast nodes.

There is also thread-level parallelism that uses multicore processors to reduce power consumption and improve latency between cores. Figure 3 shows the classification and measurement process of marketing forecasts.

Each class is a set of objects determined by a subset of dimensions, and different classes can be represented by different subsets of dimensions, so there is a subspace cluster that finds classification not in the entire data space but in the subspace of the data. The subspace clustering is determined by the membership degree of the object to the class and the subspace of each class. The starting point is to consider the correlation between feature words and categories. The greater the correlation between the feature words and a certain category is, the more they should be retained as the feature words of that category. Association analysis is to find interesting associations or related connections between itemsets from a given data set. If there is a certain regularity between the values of two or more variables, it is called an association. The purpose of association analysis is to find the hidden association network in the database. Generally, two thresholds of support and credibility are used to measure the correlation between two elements. Prediction is to find the law of change from the historical data, establish a model, and use the model to predict the types and characteristics of new sample data. Prediction in the general sense refers to the use of regression methods to predict continuous or ordered values. Forecasting is concerned with accuracy and uncertainty, which are usually measured by forecast variance.

3.4. Iterative Optimization of Model Weights. Marketing node classification refers to the process of determining a category for each marketing node in the set of marketing

nodes based on predetermined subject categories. Marketing node classification is a typical machine learning problem. The purpose of classification is to allow the machine to learn a classification rule that can map Web marketing nodes to one or more existing topic categories to help users find it quickly and accurately to the marketing node needed. Marketing node classification is generally divided into two stages: training and classification. Among them, calculation and query are independent of each other and do not affect each other; its structure is helpful for parallel processing. A general criterion for a good division is as follows: objects in the same class should be “close” or related as much as possible, and objects in different classes should be “far away” or different as much as possible, even if the following criterion functions are the smallest. Figure 4 shows the iterative distribution of model weights.

The clustering feature is a triplet, which summarizes the subcluster information of the given object. The cluster feature tree is a highly balanced tree that stores the features of hierarchical clustering. A special clustering algorithm is used to cluster the leaf nodes of the CF tree. The sparse leaf nodes are divided into outliers and deleted, and dense clusters are merged into larger clusters. This stage is the macro clustering stage. An object is inserted into the latest leaf entry. If this operation destroys the balance of the CF tree, you can use a method similar to B+ tree construction to dynamically adjust the structure of the number to reach the balance again. The fuzzy set method is a method that uses fuzzy set theory to perform fuzzy evaluation, fuzzy decision-making, fuzzy pattern recognition, and fuzzy clustering analysis on practical problems, which is an effective method to deal with uncertain problems in an earlier application. The higher the complexity of the system is, the stronger the ambiguity is. The fuzzy set theory uses the degree of

FIGURE 3: Marketing forecast classification measurement process.

FIGURE 4: Iterative distribution of model weights.

membership to describe the performance of fuzzy things. These technologies can be used to examine abnormal forms of data and then use various statistical and mathematical models to interpret these data and explain the market laws and business opportunities hidden behind these data.

4. Application and Analysis of Internet Marketing Prediction Model Based on Parallel K-Means Algorithm

4.1. Parallel K-Means Algorithm Data Feature Dimensionality Reduction. In order to verify the validity of the clustering results obtained by the clustering algorithm in this paper, four common external metrics (F-measure, entropy, purity, and interactive information) are mainly used to evaluate the

clustering results. According to the two directions of hierarchical clustering, it can be divided into two categories. (1) The bottom-up aggregation method: at the beginning, each object is regarded as a separate group, and then similar categories are merged successively until the termination condition is met. (2) Self-grouping: the top-down splitting method initially treats all objects as a group and then gradually splits the major categories into smaller categories until they are divided into k groups or other termination conditions are reached. Data set: we select the training node set D and the test node set T . D contains six types of nodes with a total of 150 articles, and T contains six types of 50 articles for a total of 300 nodes. Noise set: we select a total of 20 nodes that do not belong to any of the above six categories to form a noise set N . Network structure: 150 nodes in the input layer. The scale of the output layer is 10×10 , 15×15 , 20×20 . Figure 5 is the data partition of the parallel K-means algorithm.

We select n nodes from the noise set N and add them to the training set D . The ratio of the number of noise to the number of original training samples is $x\%$ ($n/150$). We train the SOM network with the same configuration and also gather the data into 6 kinds. Then, we use the t -test to count the proportion of misclassified marketing nodes. The experimental results can be seen for SVM method and T-SVM. The SVM method is very similar in various performance indicators. The performance values of the three are basically close in terms of macro average and micro average, which are our most commonly used indicators for evaluating the performance of classifiers. The indications fluctuate between 94% and 95.5%, and the difference is not large. This proves that the support vector preselection method based on pairwise clustering with known category information can significantly reduce the classifier's performance on the one hand. Its advantage is that there is no need to determine the number of target clusters in advance, it has good noise resistance, and at the same time, it can also be filtered. Training time, on the other hand, basically has no adverse effect on the accuracy of the final support vector machine, indicating that this method has certain advantages.

FIGURE 5: Parallel K-means algorithm data division.

4.2. Internet Marketing Prediction Model Simulation. In order to verify and evaluate the performance of the algorithm proposed in this paper, this paper builds a system platform for cluster analysis of marketing nodes. The system is mainly composed of three modules: preprocessing module, vector space model representation module, and clustering module. The preprocessing module is mainly responsible for preprocessing the marketing nodes. The vector space model representation module is mainly responsible for representing the nodes in a form that can be recognized by the computer. The clustering module clusters the processed marketing nodes. The realization of the system is mainly divided into the following three modules: preprocessing module, vector space model representation module, and clustering module. The preprocessing module also includes the following submodules: word segmentation, stop word removal, word frequency statistics, and feature selection. The three modules correspond to input pre-processed marketing nodes, SVM models, and clustering results. The experimental environment of this experiment is AMD4000+, 2G RAM, Windows XP, VC++ 6.0. The results of the Internet marketing forecast performance evaluation are shown in Figure 6.

It can be seen intuitively that the value of the CBKM algorithm is higher than the other two algorithms in terms of mutual information, purity, and F metric as the metric. The BKM algorithm has higher values on these three metrics than the K-means algorithm, which shows that CBKM has the best clustering performance on the selected test set, BKM has the second, and KM algorithm has the worst. The performance of the CBKM algorithm has been improved because it uses the KM algorithm and the BKM algorithm clustering algorithm to intersect the clustering results generated by these two algorithms to generate clustering subsets and recombine these subsets into new ones. Figure 7 shows the normalized distribution of the marketing threshold node set.

Randomly we select 9000 training nodes in the node library for training. These training nodes involve 5 categories: environment, computer, education, economy, and sports. Randomly we select 3000 test nodes in the remaining part of the node library for testing. After scanning the word frequency of the marketing node, we still need to do a very important work, which is to remove the stop words. The removal of stop words in marketing nodes is usually taken as a preprocessing operation in marketing node processing

FIGURE 6: Evaluation of Internet marketing forecast performance.

FIGURE 7: Normalized distribution of marketing threshold node set.

(including marketing node classification, marketing node filtering, and marketing node summary). If it is included in the stop vocabulary, we delete the word from the marketing node thesaurus and then take the word. The next word in the set continues to be compared with the stop word list until the word in the marketing node is processed. For a specific sample, only the kernel SOM algorithm corresponding to certain kernel functions can show its superiority.

4.3. Case Application and Analysis. We first use the word segmentation software ICTCLAS to process the word segmentation of marketing nodes and then scan the entire marketing node according to the segmented nodes, that is, counting the number of times that all nodes appear in the marketing node, which is the feature of the training marketing node. In this experiment, the word frequency statistical processing of marketing nodes is realized through the C++ language, and the statistical results are stored in the Access data table. Given a database containing n objects or original groups, partitioning methods construct k partitions of the data, each partition represents a cluster, and $k \leq n$. Its advantage is that there is no need to determine the number

FIGURE 8: Vector data clustering curve of marketing nodes.

FIGURE 9: The density distribution of the vector information value of the marketing node.

of target clusters in advance, it has good noise resistance, and at the same time, it can also be filtered. In other words, it divides the data into k groups while meeting the following requirements: each group contains at least one object, and each object must belong to only one group. This method is very suitable for discovering spherical clusters. In order to cluster large-scale data and deal with clusters of complex shapes, the partition-based method needs to be further expanded. Figure 8 is the vector data clustering curve of marketing nodes.

Since the vector of each marketing node is rewritten after the TF-IDF method processes the weight of each feature word, the weight of each dimension of the vector representing the marketing node is greater than zero and the modulus of the vector is equal to 1. All marketing node vectors in the node set fall on a unit hypersphere with a radius of l , only falling on the part of the hypersphere where the values of all dimensions are positive. Simplified to the simplest case, if the dimension of the vector is 2, then all marketing node vectors fall only on the arc of the first quadrant of the unit circle, and the distribution results are as written. Therefore, this article normalizes the marketing node vector, on this basis making the angle cosine measure and the Euclidean distance measure equivalent so that after the marketing node normalization processing, the density radius of a certain marketing node vector is determined. It is convenient to use the binary search method. Figure 9 is the density distribution of the vector information value of the marketing node.

First, we calculate the mutual information value of each node, then arrange the entries in descending order of their

mutual information value, and extract the set of nodes whose mutual information value is greater than a certain threshold as the feature word set. By extracting the features of the training marketing node set in this way, we get 1000 feature nodes to represent the vector space of the training marketing node set used in this article. The first column is the feature vector of the marketing node we extracted based on the mutual information value. The second column is the frequency of the feature vector in the current marketing node. The third column is the word frequency of the feature vector in the entire training set and records the mutual information value of the feature words. The fourth column is the normalized IDF value representation of the feature words contained in the current marketing node. The fifth column corresponds to a vector (i.e., the vector representation of the current marketing node). It is the input part of the marketing node when the marketing node is classified. Using the pairwise clustering preprocessing method based on the known category information proposed in the fourth part of this paper, the time overhead in the cluster preprocessing stage is relatively large (it takes 394.4 seconds to perform the cluster preprocessing stage), so that the samples involved in training support vector machines are greatly reduced compared with the two methods of SVM and T-SVM, so that the time cost of the support vector machine training phase is much smaller than the time cost of SVM and T-SVM training support vector machine. Therefore, the total time cost of training the classifier in the B-SVM method is the smallest.

5. Conclusion

Based on the parallel K-means algorithm, this paper proposes a solution for clustering large-scale data on the Internet distributed computing platform. The Internet shields the underlying implementation details of distributed computing, making it easy for users to deploy Internet platforms. At the same time, due to the use of technical means such as multinode redundant backup and automatic load balancing, the Internet has low requirements on the hardware environment and has a bright future for development. The mean clustering algorithm makes full use of the known conditions of the training marketing node category information and reduces the nonsupport vectors in the training samples as much as possible in the training sample preprocessing stage in the way of pairwise clustering so that they can finally participate in the training support. The number of samples of the vector machine is greatly reduced, thus generating a marketing node classification model that has the classification accuracy equivalent to the traditional support vector machine classification method and has a greater improvement in time complexity. At the same time, under the Internet platform, the traditional K-means algorithm is reprogrammed according to the MapReduce programming model, and the Canopy algorithm MapReduce is paralleled. It is possible to complete the clustering of marketing nodes on a large-scale data set on a cheap platform. This paper also implements an improved support vector machine classification of marketing nodes. Because the selection of the

penalty factor and kernel function parameters of the support vector machine has a great influence on the classification effect, this article first uses the particle swarm optimization algorithm to optimize the penalty factor and the kernel function parameters and then uses the support vector machine to classify the effectiveness of support vector machine classification.

Data Availability

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

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

References

- [1] S. Kumari, N. Patil, P. Nankar, and M. Kulkarni, "CUDA parallel computing framework for stock market prediction using K-means clustering," in *Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 467–473, Trichy, India, September 2020.
- [2] Y. Li, R. Yang, and P. Guo, "Spark-based parallel OS-ELM algorithm application for short-term load forecasting for massive user data," *Electric Power Components and Systems*, vol. 48, no. 6-7, pp. 603–614, 2020.
- [3] O. A. Ghoneim, "Traffic jams detection and congestion avoidance in smart city using parallel K-means clustering algorithm," in *Proceedings of the Cognition and Recognition*, pp. 21–30, Singapore, September 2018.
- [4] X. Dong, S. Deng, and D. Wang, "A short-term power load forecasting method based on K-means and SVM," *Journal of Ambient Intelligence and Humanized Computing*, pp. 13–15, 2021.
- [5] O. Koren, C. A. Hallin, N. Perel, and D. Bendet, "Decision-making enhancement in a big data environment: application of the K-means algorithm to mixed data," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, 2019.
- [6] Y. Xu, C. Yang, S. Peng, and Y. Nojima, "A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning," *Applied Intelligence*, vol. 50, no. 11, pp. 3852–3867, 2020.
- [7] O. Koren, C. A. Hallin, N. Perel, and D. Bendet, "Enhancement of the k-means algorithm for mixed data in big data platforms," in *Proceedings of the Intelligent Systems Conference*, pp. 1025–1040, London, UK, September 2018.
- [8] X. Lv, X. Cheng, and Y. Tang, "Short-term power load forecasting based on balanced KNN," *Materials Science and Engineering*, vol. 322, no. 7, Article ID 072058, 2018.
- [9] Z. Fang and C. Chiao, "Research on prediction and recommendation of financial stocks based on K-means clustering algorithm optimization," *Journal of Computational Methods in Science and Engineering*, vol. 21, pp. 6–9, 2021.
- [10] R. Tripathy and R. K. Nayak, "Eukaryotic plasma cholesterol prediction from human GPCRs using K-means with support vector machine," *Advanced Deep Learning for Engineers and Scientists*, pp. 243–257, 2021.
- [11] L. Li, J. Wang, and X. Li, "Efficiency analysis of machine learning intelligent investment based on K-means algorithm," *Ieee Access*, vol. 8, pp. 147463–147470, 2020.
- [12] N. Wei, L. Yin, C. Li, C. Li, C. Chan, and F. Zeng, "Forecasting the daily natural gas consumption with an accurate white-box model," *Energy*, vol. 232, Article ID 121036, 2021.
- [13] M. Ghayekhloo, R. Azimi, M. Ghofrani, M. B. Menhaj, and E. Shekari, "A combination approach based on a novel data clustering method and Bayesian recurrent neural network for day-ahead price forecasting of electricity markets," *Electric Power Systems Research*, vol. 168, pp. 184–199, 2019.
- [14] H. Pan, Y. Lei, and S. Yin, "K-means clustering algorithm for data distribution in cloud computing environment," *International Journal of Grid and Utility Computing*, vol. 12, no. 3, pp. 322–331, 2021.
- [15] Y. M. Mao, D. J. Gan, D. S. Mwakapesa, Y. A. Nanehkaran, T. Tao, and X. Yu Huang, "A MapReduce-based K-means clustering algorithm," *The Journal of Supercomputing*, pp. 21–22, 2021.
- [16] A. G. Veerasha, H. A. M. Prasanna, M. V. L. Kumar, and T. Ananthapadmanabha, "An adaptive power system management with DG placement and cluster-based load forecasting by CS, K-means and ANN algorithms," *International Journal of Power Electronics*, vol. 13, no. 3, pp. 380–398, 2021.
- [17] J. Huang, Y. Liang, H. Bian, and X. Wang, "Using cluster analysis and least square support vector machine to predicting power demand for the next-day," *IEEE Access*, vol. 7, pp. 82681–82692, 2019.
- [18] H. Omrani, B. Parmentier, M. Helbich, and B. C. Pijanowski, "The land transformation model-cluster framework: applying K-means and the Spark computing environment for large scale land change analytics," *Environmental Modelling & Software*, vol. 111, pp. 182–191, 2019.
- [19] W. XingFen, Y. Xiangbin, and M. Yangchun, "Research on user consumption behavior prediction based on improved XGBoost algorithm," in *Proceedings of the IEEE International Conference on Big Data*, pp. 4169–4175, Seattle, WA, USA, December 2018.
- [20] J. Li, S. Pan, and L. Huang, "A machine learning based method for customer behavior prediction," *Tehnički Vjesnik*, vol. 26, no. 6, pp. 1670–1676, 2019.
- [21] A. Onan and S. Korukoglu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, vol. 43, no. 1, pp. 25–38, 2017.
- [22] H. Li, R. Liu, J. Wang, and Q. Wu, "An enhanced and efficient clustering algorithm for large data using map reduce," *IAENG International Journal of Computer Science*, vol. 46, no. 1, pp. 61–67, 2019.
- [23] K. Song, X. Yao, F. Nie, X. Li, and M. Xu, "Weighted bilateral K-means algorithm for fast co-clustering and fast spectral clustering," *Pattern Recognition*, vol. 109, Article ID 107560, 2020.