Proceedings of the 3rd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/71/20241485

Application of K-Means Algorithm in Marketing

Qiyan Chen[1,a,*]

[1] *Wuhan University of Science and Technology, Wuhan, Hubei province,430081, China*
a. 2398352718@qq.com
**corresponding author*

Abstract: The purpose of this paper is to explore the application of K-means cluster analysis in the field of marketing. Firstly, this paper briefly introduces the research background, the research topic, and the significance of the research. K-means is an important algorithm in data mining, so this paper first gives a brief overview of data mining and introduces its main tasks. Then it focuses on explaining the related knowledge of the K-means algorithm, which includes the specific principles, advantages, improvements, and dissimilarity calculation of the K-means algorithm. In the specific combination of marketing, how to use K-means for clustering is described in detail using consumers as an example. Using the literature research method, this paper concludes that K-means cluster analysis is a very useful tool in marketing, which can help companies better understand their markets and customers, and develop more accurate marketing strategies.

Keywords: K-means algorithm, marketing, customer segmentation, data mining, clustering algorithm

1. Introduction

With the increasing competition in the market, how to effectively develop and implement marketing strategies has become the key to business success. Traditional marketing methods are often based on experience and intuition, making it difficult to achieve accurate target marketing. In recent years, data-driven marketing strategies have gradually received attention from enterprises, among which, K-means cluster analysis, as a powerful data analysis tool, provides new perspectives and methods for marketing.

This paper first introduces the principle and basic process of K-means cluster analysis and then elaborates in detail how to apply K-means to consumer clustering from the aspects of data preprocessing, feature selection, model training and result analysis. Through experiments, it can be found that K-means cluster analysis can effectively improve the accuracy and efficiency of operators' marketing strategies.

K-means cluster analysis is an unsupervised machine learning method that discovers patterns and structures in data by dividing it into multiple similar clusters. In the field of marketing, K-means cluster analysis can be used to identify and analyse potential customer groups, predict consumer behaviour, and grasp market trends. However, there are still many challenges in how to effectively use K-means cluster analysis to optimise marketing strategies.

With the increasingly fierce competition in the market, enterprises have more and more urgent needs for the refinement of consumer groups, and K-means clustering analysis, as an effective data

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

classification method, can help enterprises to understand the consumer groups in-depth, and then formulate more targeted marketing strategies. This paper will discuss the application of K-means cluster analysis in marketing and its research significance.

The combination of K-means cluster analysis and marketing strategy has a wide range of application prospects and research significance. Through the refined division of consumer groups, enterprises can formulate more targeted marketing strategies to improve market competitiveness. However, K-means cluster analysis still has certain limitations, such as the sensitivity to the selection of the initial cluster centre. This paper will also look at the future research direction can include improving the K-means algorithm, as well as combining it with other advanced techniques and methods to achieve a more stable and efficient consumer group segmentation for future reference in related fields [1].

2. Data Mining

2.1. Overview of Data Mining

Data mining is the automatic extraction of implicit, unknown models and rules with application value from huge amounts of data. Data mining is the process of discovering useful information automatically in large stored databases. Data mining is an important part of Knowledge Discovery in Databases (KDD). So, what is KDD? KDD is the process of converting unprocessed data into useful information. The process includes everything from inputting data to data preprocessing to data mining, to post-processing, and finally to information. Data mining brings together knowledge from many disciplines; it requires sampling, estimation and hypothesis testing from statistics, as well as search algorithms, modelling techniques and learning theories that incorporate artificial intelligence, pattern recognition and machine learning [2]. Data mining can be seen as the intersection of machine learning and databases, and it primarily uses techniques provided by the machine learning community to analyse large amounts of data and techniques provided by the database community to manage large amounts of data [3].

2.2. Tasks of Data Mining

Usually, data mining tasks are divided into two broad categories.

Prediction tasks. Predict the value of a specific attribute (target variable) based on the value of the attribute (description variable). For example, predicting house prices based on house type, location, etc.

Descriptive tasks. Finding patterns of potential connections in the data, such as correlations, trends, clusters, anomalies, etc.

This paper focuses on four main data mining tasks.

(1) Predictive modelling: This refers to modelling the target variable in a way that describes the function of the variable. There are two types of predictive modelling tasks: classification, which predicts discrete target variables; and prediction, which predicts continuous target variables.

(2) Correlation analysis: discovers patterns of strongly correlated features in data.

(3) Cluster analysis: to discover clusters of closely related observations.

(4) Anomaly analysis: identifying observations whose characteristics are significantly different from those of other data.

3. Cluster Analysis

3.1. Overview of Cluster Analysis

Clustering is the process of dividing data objects into classes and clusters based on some characteristic criteria that make the similarity of data objects within the same cluster as large as possible [4].

Types of clustering:

(1) Divisional clustering: divides data objects into non-overlapping subsets (clusters).

(2) Hierarchical clustering: Allow clusters to contain subsets within them, and form all clusters into a tree. Hierarchical clustering can be viewed as a sequence of division clusters.

Different types of clusters:

(1) Prototype-based: clusters are collections of objects that tend to be spherical. The distance of each object to the prototype (cluster centre) of the cluster is less than the distance to the prototypes of other clusters. Representative algorithms are K-mean, K-modes

(2) Density-based: Clusters are dense regions of objects surrounded by low-density regions. Representative algorithms are DBSCAN, and optics.

3.2. K-means Clustering

The k-means clustering algorithm is a more commonly used algorithm in cluster analysis because of its simplicity and scalability [5]. Kmeans clustering is capable of automatic clustering when dealing with unlabelled datasets and can be applied to different types of datasets, and thus belongs to unsupervised learning. In addition, K-means clustering is also very interpretable and can clearly show the results of clustering, which helps people to better understand the structure and distribution of the dataset.

3.2.1. History of K-means Clustering

The Kmeans clustering algorithm was first proposed in 1957 by Stuart Lloyd as part of the pulse code modulation technique. However, the algorithm was not published until 1965 by E.W. Forgy and was named "K-means" in 1967 by James MacQueen in his paper "Some Methods for the Classification and Analysis of Multivariate Observations".

3.2.2. K-means Clustering Algorithm

Step 1: Initialise the clustering centres. Randomly select k clustering centres (k needs to be determined in advance). Step 2: Assign samples to clustering centres. Calculate the distance from each sample to each clustering centre and assign each sample to the clustering centre closest to it. Step 3: Move the clustering centre. Calculate the new cluster centre such that the new cluster centre moves to the mean of all samples in this cluster. Step 4: Repeat steps 2 and 3 until the cluster centre value is no longer updated, or the SSE is less than a given threshold, or the maximum number of iterations is reached and the algorithm stops [6].

3.2.3. Features of K-means Clustering

Advantages: the algorithm is relatively scalable and has high efficiency when dealing with large data sets.

Application limitations: (1) the number of clustering clusters needs to be specified in advance; (2) it is only applicable to numerical attribute clustering; and (3) it is sensitive to noisy and anomalous data;

3.2.4. Improvements in K-means Clustering

Intuitive selection: Determine the k value based on the results of visualisation.

Elbow method: obtain multiple models with different k, respectively, then calculate the corresponding SSE and plot the SSE-k curve to observe and obtain the k value. The k-value corresponding to the elbow is closest to the true number of clusters.

Contour coefficient method: find the value of k that maximises the contour coefficient s. s is the contour coefficient of the clustering result, and the larger s is, the more reasonable and effective the clustering is. The contour coefficient s and the sum of squared errors SSE are both measures of whether the clustering is reasonable or not. SSE evaluates the error within clusters. s evaluates the error within clusters and the distance between clusters.

3.2.5. Calculation of the Degree of Dissimilarity

K-means uses a distance measure of the dissimilarity of the samples and is suitable for continuous type attributes. Depending on the problem and the type of attribute, the appropriate proximity measure is selected. Documents can use Yuhuan similarity; coding uses Hamming distance; transactional data uses the Jaccard coefficient, etc.

4. Application of K-means Cluster Analysis in Marketing

Marketing is the process by which an enterprise analyses the market and customer needs and achieves the sales and promotion of goods and services through appropriate products, prices, channels, marketing strategies and other means, so as to obtain commercial benefits.

The purpose of marketing is to satisfy customers' needs, increase sales and market share, and enhance corporate brand awareness and reputation, so as to obtain more business opportunities and profits.

In marketing, K-means can be used to segment consumers, select target markets, and develop marketing strategies.

The specific steps of consumer segmentation, for example, are as follows:

4.1. Data Collection and Data-Processing

When collecting data, it is important to note that the data collected needs to reflect consumer characteristics, which can be age, income, gender, geographic location, consumer behaviour, buying preferences, etc. This can be obtained from databases, questionnaires, data crawlers, etc. The collected data is then subjected to data pre-processing.

4.1.1. Introduction to Data Pre-processing

Data preprocessing is the process of aggregating, sampling, cleaning, transforming and normalising data and using it directly for analytical modelling. There is a close connection between data mining and data preprocessing, the better the quality of data, the more accurate the results of data mining. Especially when mining noisy data, inconsistent data, and incomplete data, data preprocessing plays an important role [7].

4.1.2. Main Tasks of Data Pre-Processing

Aggregation. Aggregation is the integration of two or more datasets into a single dataset. The goal is to merge data. The advantages are that the data can be viewed from a more advanced view and is relatively less computationally expensive, and the behaviour of groups of objects or attributes is usually more stable than that of individual objects [8]. The disadvantage is that details may be lost and attention cannot be paid to the detailed characteristics of individuals.

Sampling: Sampling is a common method of selecting a subset of data objects for analysis. The purpose of sampling is to resolve unbalanced datasets or to address datasets that are super-sized and too computationally expensive.

Data Cleaning: Data cleaning mainly includes processing missing values, outliers, de-emphasis and other operations.

Feature Transformation: For a certain feature of the original data, use the appropriate method to transform the distribution, scale, etc. of the data in order to meet the needs of modelling on the data. Feature transformation includes function transformation and discretization. It can solve the problems of data scale and value range.

Dimensional approximation: It can solve the problems of dimensional catastrophe and feature insignificance. What is dimensional disaster? Dimensional catastrophe means that with the increase in data dimensions, many data analyses will become very difficult, and the data will become more and more sparse in the space it occupies, i.e., for high-dimensional data, the accuracy of classification will be reduced, and the quality of clustering will be reduced.

4.2. Feature Engineering

Feature extraction and selection of consumer data, converted into feature vectors suitable for the K-means clustering algorithm, can be performed based on the distribution of the data and the correlation between the features to select more discriminating features [9].

4.3. K-Means Algorithm

Based on the selected feature vectors, consumers are segmented using the K-means clustering algorithm. The k-value (number of clusters) can be determined using the profile coefficient method or the elbow method mentioned above. During the clustering process, attention needs to be paid to the initialisation and convergence conditions in order to avoid generating locally optimal solutions.

4.4. Analysis of Results

The clustering results are analysed, including the centre of the clusters, the characteristics of each cluster, and the number of consumers in each cluster. Consumers can be classified and named according to the clustering results, such as "young white-collar workers", "middle-aged families", "elderly health care" and so on.

4.5. Marketing Strategy Development

Develop personalised marketing strategies based on clustering results and consumer characteristics. Different product positioning, pricing strategies, promotion methods, etc. can be developed based on consumer preferences and purchasing behaviours of different clusters [10].

4.6. Modeling Assessment

Use existing marketing strategy data for model evaluation, including metrics such as accuracy, recall, and F1 value. Adjustments and optimisations can be made based on the evaluation results to improve the accuracy and reliability of the model.

5. Research Challenges and Perspectives

5.1. Challenges in User Quality Prediction Research

In many practical applications, the data to be processed often contain both numerical and sub-typical attributes, at this time the use of K-means or K-modes algorithm cannot effectively solve the problem [11]. At the same time the face of dealing with high-dimensional data will encounter serious dimensionality disaster problems, the higher the data dimensionality, the worse the clustering effect. The K-means algorithm is sensitive to the initial clustering centre, the choice of a different clustering centre will produce different clustering results and different accuracy, the practice of randomly selecting the initial clustering centre will lead to algorithmic instability, and there is a possibility of falling into the situation of the local optimum.

5.2. K-Means Future Development Trend and Outlook

5.2.1. Integration with Other Technologies

The K-means algorithm can be combined with other clustering algorithms or machine learning algorithms to form more effective algorithms. For example, the K-means algorithm can be combined with machine algorithms to further improve the accuracy of user quality prediction, or with correlation analysis to discover more user group characteristics.

5.2.2. Processing of High Latitude Data

The k-means algorithm is not effective in handling high dimensional data, but there are many applications in high dimensional data, such as in medicine, social media, image processing, etc. Therefore, in future research, this paper can explore how to deal with high dimensional data more effectively so that K-means can be applied to more scenarios [12].

5.2.3. Evaluation of Clustering Effects

Clustering algorithms are currently widely used in various domains, but the existing evaluation metrics are mainly for datasets in general domains. Future research can explore how to evaluate clustering effects on domain-specific datasets to improve the effectiveness of clustering algorithms in specific domains. Now there are existing evaluation metrics that can measure the clustering effect to a certain extent, and new evaluation metrics can be explored in future research development to evaluate the clustering effect more comprehensively and correctly.

6. Conclusion

In this paper, this paper explores the application of K-means cluster analysis in marketing strategy optimization. By applying K-means cluster analysis method to marketing strategy research, this paper is able to better understand customer needs and formulate more accurate target marketing strategies, thus improving marketing effectiveness. First, this paper provides an overview of the principles and applications of K-means cluster analysis in the field of marketing. K-means cluster analysis is an unsupervised machine learning method that discovers patterns and structures in data by dividing the data into multiple similar clusters. In the field of marketing, K-means cluster analysis can be used to identify and analyse potential customer groups, predict consumer behaviour, and capture market trends [13]. This method can help companies better understand customer needs, identify market opportunities, and improve marketing effectiveness.

Next, this paper clarified the research purpose of this paper, which is to explore how K-means cluster analysis can optimise marketing strategies. In the research process, this paper adopted the K-means cluster analysis method, which includes the steps of data collection, data processing and cluster analysis. Data collection involves obtaining data from sources such as market research and user behaviour; data processing includes steps such as data cleaning and feature extraction; and cluster analysis divides the data into different clusters to discover potential customer groups. These steps are important for the final clustering results and the development of marketing strategies.

Finally, this paper summarizes the findings of this paper. By applying K-means cluster analysis to marketing strategy research, this paper is able to better understand customer needs and develop more accurate target marketing strategies. At the same time, this paper also proposes some methods and future research directions for optimising K-means cluster analysis to provide reference and guidance for work in related fields.

In conclusion, K-means cluster analysis has a wide range of application prospects in marketing strategy optimisation. Through continuous in-depth research and practical application, this paper can better grasp the market demand and formulate more accurate marketing strategies, so as to create greater value for the enterprise.

References

[1] *M.J.A Berry and G.Linoff. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.Wiley Computer Publishing,2nd edition,2004.*

[2] *Pang-Ning Tan,Michael Steinbach,Vipin Kumar.Introduction to Data Mining.People's Posts and Telecommunications Publishing House,2011:2-6.*

[3] *R.Roiger and M.Geatz.Data Mining: A Tutorial Based Primer.Addison-Wesley,2002.*

[4] *M.S.Aldenderfer and R.K.Blashfield.Cluster Analysis.Sage Publications, Los Angeles,1985.*

[5] *Huchang.Design of a User Behaviour Analysis System.[Master's Thesis, Hubei University of Technology].Wuhan: Hubei University of Technology,2011.*

[6] *P.Berkhin.Survey Of Clustering Data Mining Techniques.Technical report, Accrue Software, San Jose,CA,2002.*

[7] *V.Cherkassky and F.Mulier. Learning from Data: Concepts, Theory,and Methods.Wiley Interscience,1998.*

[8] *YAN BAI, XIAO SU, BHARAT BHARGAVA. Adaptive voice spam control with user behavior analysis. Proc. 11th IEEE Int'l Conf. on High-Performance Computing and Communications, 2009, 354-361*

[9] *J.MacQueen.Some methods for classification and analysis of multivariate observations.In Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, pages 281-297.University of California Press,1967.*

[10] *Yang Baozhen. Innovation of corporate marketing strategy[J]. Enterprise Economy,2011,30(05):76-78.*

[11] *YANG Shanlin,LI Yongsen,HU Xiaoxuan et al.Study on K-value optimisation problem in K-MEANS algorithm[J]. Systems Engineering Theory and Practice,2006(02):97-101.*

[12] *Y. Zhao and G.Karpis.Empirical and theoretical comparisons of selected criterion functions for document clustering.Machine Learning,55(3):-331,2004.*

[13] *WANG Qian, WANG Cheng, FENG Zhenyuan et al. A review of research on K-means clustering algorithm[J]. Electronic Design Engineering,2012,20(07).*