Ministry of Education of the Republic of Belarus
Educational Institution
Belarusian State University
of Informatics and Radioelectronics

UDC 004.93’14

Zhang Hengrui

MODIFIED K-MEANS ALGORITHM

ABSTRACT
for a master's degree

Speciality 7-06 0612 03 Information management systems

____

Academic Supervisor
German Yuliya Olegovna
Candidate of Technical Sciences,
Associate Professor

____

Minsk 2025

INTRODUCTION

In the era of rapid digital development nowadays, data has undoubtedly become one of the core driving forces in various fields of society. With the in-depth popularization of information technology, the Internet of Things and the Internet, the amount of data worldwide has shown an explosive growth trend, with global data volume projected to exceed 180 zettabytes by 2025. Whether it is the massive business data accumulated during enterprise operations, such as the hundreds of millions of daily transaction records on e-commerce platforms and the vast amount of real-time cargo transportation information tracked by logistics enterprises, or the huge amounts of data collected through various complex experiments and observations in the scientific research field, we are all immersed in a sea of data.

The innovation of computer technology has made data collection, storage, and transmission more convenient and efficient, which directly leads to a sharp increase in the number of devices and users that generate data. The number of global Internet users has long exceeded billions, and there are billions of dynamic updates on social platforms every day. The widespread use of mobile devices enables data to be produced anytime and anywhere.

Facing such a surging data flood, enterprises and organizations have unprecedented challenges. On the one hand, a large amount of unorganized data is like disorderly raw materials piled up in a warehouse and cannot directly provide effective support for decision-making. For example, a multinational chain enterprise wants to understand the preference differences of consumers in different regions for its various products to formulate targeted marketing strategies. However, the massive amount of consumer evaluation data collected from stores distributed around the world, online sales platforms, and after-sales feedback channels makes them at a loss and difficult to quickly and accurately extract valuable information. On the other hand, for scientific researchers, how to mine hidden laws and discover new knowledge from complex experimental data is also a thorny problem.

Cluster analysis technology emerged as the times require and has become a powerful weapon to solve this dilemma. It aims to divide a large-scale dataset into several relatively small subsets with high internal similarity and large differences between each other, that is, what we call “clusters”, based on the similarities or differences among data objects. In this way, the originally chaotic and disorderly data can be sorted into organized and characteristic groups, enabling enterprises to clearly understand the segmented characteristics of consumer groups and scientific researchers to capture key patterns and trends from complex data.

In the commercial field, cluster analysis has extensive and crucial application scenarios. Market segmentation is an important part of it. Enterprises use cluster analysis to accurately divide different target customer groups based on data in many dimensions such as consumers’ age, gender, consumption habits, and purchase frequency. For example, a beauty brand can analyze the historical data of consumers’ purchases of beauty products and cluster consumers into different groups. Then, it can design unique product packaging, formulate differentiated promotion strategies, and develop new products that meet the needs for each group to achieve precise marketing and maximize the input-output ratio of marketing resources.

In the scientific research category, cluster analysis also plays an indispensable role. In the field of biology, researchers use cluster analysis to process the gene sequence data of biological species and cluster similar gene sequences together, which helps to discover new species phylogenetic relationships and explore the mysteries of species evolution. In astronomy, facing the massive celestial body data observed in the vast universe, cluster analysis can classify celestial bodies with similar characteristics, assist astronomers in identifying galaxy clusters, stellar populations, and other different celestial structures, and promote a deeper understanding of the laws of the universe’s evolution.

However, although clustering analysis has shown great potential for application, there are still many challenges that need to be overcome in the field. Different types of data have their own unique distribution characteristics, noise interference, and dimensionality catastrophe, which require clustering algorithms to be highly adaptable and accurate. For example, when dealing with high-dimensional sparse data, traditional clustering algorithms may fall into the local optimal solution, which cannot effectively find the real inner structure of the data; in the face of a large amount of noise in real data, the algorithm's stability and robustness face a severe test. Therefore, it is of great practical significance and scientific value to continue to study clustering analysis technology and develop more advanced and efficient clustering algorithms. The master's thesis focuses on this, is committed to in-depth investigation of the key technologies and methods in cluster analysis. The subsequent chapters will consider various K-means algorithms, cluster construction, selection of metrics, and other core points one by one to carry out a detailed discussion. The purpose of this master's thesis is to develop an algorithm to improve standard K-means.

GENERAL CHARACTERISTICS OF THE WORK

The purpose of this research is to modify the traditional K-means algorithm to have higher accuracy and adapt it to handle dynamic data.

The research is relevant due to more and more information needs to be categorized in today's information age. Standard k-means has disadvantages such as unstable clustering results and inability to handle dynamic data. However, a large amount of data needs to be categorized more accurately.

The object of the study is clustering and different approaches to it solving. The subject of the research is variances of K-means algorithms, evaluation metrics, KDTree, annealing approach.

For the purposes of this research, it is assumed that the following tasks will be solved:

– Analyzing different K-means algorithms and algorithm evaluation metrics. This stage includes researching relevant scientific papers and available technical solutions for the implementation of cluster analysis algorithms and metrics for algorithm evaluation.

– Modification of the cluster analysis algorithm. The task at this stage is to modify the K-means algorithm based on the results of the analysis.

– Implementation of the algorithm. At this stage, it is necessary not only to develop the algorithm, but also to ensure its correct implementation in the form of program code.

– Implementation and usage.

All of the above tasks are designed to achieve the main goal of this research, which is to modify the standard K-means algorithm to solve some of the shortcomings of standard K-means.

Provisions for defense:

1. Using KDTree and greedy algorithms to get the initial center of mass. This gives higher efficiency in initial center of mass selection as well as better results than the normal greedy algorithm.

2. Annealing incremental assignment of samples with combining a physical simulation of the annealing process ensures higher final accuracy however with greater time to get solution.

The topic of the dissertation work corresponds to the list of priority areas determined by the Decree of the President of the Republic of Belarus dated 07.05.2020 (No. 156) “On priority areas of scientific and technical activity in the Republic of Belarus for 2021-2025” (direction of development of the information society, electronic state and digital economy).

The main provisions and results of the research were discussed at various conferences:

Proceedings of the 60th scientific conference of graduate students, undergraduates and students, BSUIR, Minsk – 2024;

Information Technologies and Systems 2024 (ITS 2024): Proceedings of the international scientific conference, Minsk, November 20, 2024, BSUIR;

Internauka: Electronic Scientific Journal. Part 1. – 2025. – № 13(377).

SUMMARY OF THE WORK

The focus of this work is to research and modify the standard K-means algorithm in order to maintain stable and good accuracy in cluster analysis.

The focus of this research is to propose modifications corresponding to the various shortcomings in the standard K-means algorithm and to analyze the clustering results of the new algorithm in detail.

The initialized clustering centers are optimized using greedy algorithm and KDTree. The algorithm takes into account the various uncertainties of random initialization, using the greedy algorithm to find out the better clustering centers and KDTree to speed up the search. This makes the algorithm to be greatly improved in terms of accuracy.

Clusters are assigned using an annealing incremental algorithm. The algorithm considers incremental datasets and solves the error accumulation problem of incremental K-means using simulated annealing, which greatly improves the accuracy and allows the algorithm to handle dynamic datasets.

Finally, experiments were conducted to perform cluster analysis using randomly generated clusters and actual mall customer data, considering different scenarios with different dimensions, different dataset sizes, and dynamic dataset processing, and all of them yielded excellent results.

CONCLUSION

With the explosive growth of data size and increasing complexity, traditional clustering algorithms face significant challenges in dealing with high-dimensional, dynamic and unstructured data. In this research, an improved algorithm incorporating greedy initialization strategy and annealing incremental optimization is proposed to address the problems of initial center of mass sensitivity, local optimal traps and insufficient dynamic adaptability of the classical K-means algorithm. The geometric rationality of the initial center of mass is significantly improved by the density-aware center of mass replacement mechanism and KDTree-accelerated neighborhood search; combined with the temperature-modulated soft allocation strategy, the algorithm achieves a balance between global exploration and local optimization in the dynamic data stream. Experimental results show that the improved algorithm outperforms the traditional K-means and its variants in internal and external evaluation metrics such as silhouette coefficient, V-measure, etc., and especially exhibits stronger robustness in noise interference and high-dimensional sparse data scenarios. For example, on the shopping mall customer segmentation dataset, the improved algorithm has significant improvement in each metric over the rest of the algorithms, which verifies its practical value in real business scenarios.

The theoretical contribution of this research is not only to propose an efficient initialization and optimization framework, but also to reveal the potential of annealing mechanism in cluster analysis. By introducing the idea of entropy regularization in statistical physics, the algorithm mathematically constructs a continuous and differentiable optimization space, which provides an interdisciplinary methodological reference for subsequent research. In addition, the integration of dynamic cooling rate design and incremental learning mechanism opens up new paths for real-time data stream processing, for example, promising applications in real-time monitoring of IoT devices and dynamic community discovery in social networks. Future research can further explore adaptive temperature scheduling strategies, incorporate reinforcement learning to dynamically adjust parameters, or introduce graph neural networks to capture the data streaming structure to cope with more complex multimodal distribution scenarios.

Although significant progress has been made in this research, attention still needs to be paid to the scalability of the algorithms in ultra-large-scale data. Recent research has shown that a distributed clustering structure based on edge computing can effectively alleviate computational bottlenecks, and improved algorithms can be combined with distributed architectures in the future to further enhance their engineering utility. In conclusion, this research provides an important reference for the theoretical innovation and industrial landing of clustering algorithms, and also injects new vitality into the AI-driven unsupervised learning paradigm.

LIST OF PUBLISHED WORKS

1-A. Hengrui, Z. K-means clustering algorithm and improvement methods. / Hengrui, Z, Yu. O. German // Proc. of the 60th scientific conference of graduate students, undergraduates and students / BSUIR, Minsk – 2024. – P. 39–40.

2-A. Hengrui, Z. Principles and applications of fuzzy clustering algorithms. / Hengrui, Z, Yu. O. German // Information Technologies and Systems 2024 (ITS 2024): Proc. of the international scientific conference, Minsk, November 20, 2024 / BSUIR, Minsk – 2024. – P. 196–197.

3-A. Hengrui, Z. One local language model. / German, Y. O., German, O. V., Nasr, S. N., Hengrui, Z., & Caigui, Zh. // Internauka: Electronic Scientific Journal. Part 1. – 2025. – № 13(377). – P. 30–35.