IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 7, May 2024

PPHOPCM Privacy-Preserving High-order Possibilistic C-Means Algorithm for Big Data Clustering with Cloud Computing

Vijaya Nanthini[1] and R. Mahalakshmi[2]

PG Student, Department of Computer Applications[1]
Associate Professor, Department of Computer Applications[2]
Vels Institute of Science Technology and Advanced Studies, Pallavarm, Chennai, India
vijayananthini92@gmail.com and rmahalakshmi.scs@velsuniv.ac.in

Abstract*: In image analysis and knowledge discovery, the possibilistic c-means technique (PCM), a crucial fuzzy clustering tool in data mining and pattern recognition, is widely used. Nevertheless, because PCM was first developed for small structured datasets, it might be difficult to get good clustering results for huge data, especially when the data is diverse. The research proposes a high-order PCM approach (HOPCM) for big data clustering, which resolves this problem by optimizing the objective function using tensor space. We also build a distributed HOPCM method for extraordinarily large amounts of heterogeneous data using MapReduce. Finally, we create a privacy-preserving HOPCM algorithm (PPHOPCM) to protect sensitive data on cloud servers by utilizing the BGV encryption method. PPHOPCM approximates the functions for updating the membership matrix and clustering canters as polynomial functions, facilitating the safe computation of the BGV method. Based on trial results, PPHOPCM may effectively cluster a large volume of heterogeneous data using cloud computing without revealing personal data*.

Keywords: possibilistic c-means technique.

I. INTRODUCTION

Big data is expanding rapidly along with the popularity of social media sites like Facebook and Twitter and personal computers. Large data sets are usually heterogeneous, meaning that any object Inside the set has multiple modes. Large data sets, in particular, comprise a wide Variety of interconnected object types, including audio, video, and text. This leads to a high degree of structural heterogeneity, encompassing both organized and unstructured data. Furthermore, even though several object kinds are related to one another, they each contain a different type of information. For example, a piece of sport video with meta-information uses a large number of subsequent images to display the exercise process and uses some meta-information, such as annotation and surrounding texts, to show additional information which are not displayed in the video, for instance the names of athletes. Although the subsequent images pass on different information from the surrounding texts, they describe the same objects from different perspectives. Furthermore, big data are usually of huge amounts. For example, Facebook, the famous social websites, collects about 500 terabytes (TB) data every day. These features of big data bring a challenging issue to clustering technologies.The goal of clustering is to group items with similar qualities together by dividing them into multiple groups based on unique metrics. Data engineering and knowledge discovery have successfully used clustering algorithms. Researchers and data engineers are paying close attention to big data clustering as big data grows in popularity. For instance, by expanding on their earlier image-text clustering technique, Gao et al. created a graph-based co-clustering approach for huge data. In order to cluster large data sets, Chen et al. created a nonnegative matrix tri-factorization approach that captures the correlation across the various modalities. Zhang et al. used the tensor vector space to represent the correlations over the many modalities in order to offer a high-order clustering approach for huge data. But for the next two reasons, they find it hard to cluster massive data efficiently, especially heterogeneous data. They are unable to achieve the intended findings because, in the first place, they concatenate the characteristics from several modalities linearly and neglect the intricate relationships concealed in the heterogeneous data sets. Secondly, their use is limited to

ISSN 2581-9429 IJARSCT

DOI: 10.48175/IJARSCT-18608

tiny data sets due to their high temporal complexity. They are therefore unable to effectively cluster vast volumes of heterogeneous data.

In order to address the aforementioned issues, this work suggests a high-order PCM approach for big data clustering that preserves privacy (PPHOPCM). PCM is a significant fuzzy clustering algorithm.PCM is able to prevent noise from contaminating the clustering process and effectively reflect each object's typicality to various clusters. PCM was initially created for small structured datasets, so it cannot be simply applicable to big data clustering.In particular, it is unable to capture the intricate link among the heterogeneous data object's many modalities. By extending the standard PCM technique into the tensor space, the research suggests a high-order PCM algorithm. In big data analysis and mining, tensors—also known as multidimensional arrays in mathematics—are frequently employed to represent heterogeneous data. In this research, the proposed HOPCM algorithm reveals the correlation over different modalities of the heterogeneous data object by representing each object with a tensor. We develop a distributed HOPCM algorithm based on MapReduce to use cloud servers to run the HOPCM algorithm in order to improve clustering big data efficiency. Nevertheless, when using the cloud for HOPCM, the private information usually becomes public knowledge. Consider the medical data, which is a common kind of large data. The medical records contain a great deal of sensitive information, including personal email addresses and diagnostic data. The lives and property of people will be seriously threatened by the revelation of private information. As a result, we suggest a privacy-preserving HOPCM strategy that uses the very effective BGV technique to safeguard private data on cloud servers. Despite being a completely homomorphic encryption method, BGV regrettably does not provide the division and square root operations that are required for the computation in the functions that update the membership matrix and clustering centers in the HOPCM algorithm. In order to solve this problem, we convert these functions into polynomial functions using Taylor's theorem.

II. LITERATURE SURVEY

Advances in Possibilistic C-Means Algorithm for Big Data Clustering

Introduction

In recent times, predicting share market values accurately has become increasingly challenging, often leading to significant financial losses for investors. To address these difficulties, my project focuses on developing a more reliable prediction system by analyzing large volumes of market data. This system leverages the C-Means (CM) clustering algorithm to help market experts provide better investment advice, ultimately reducing client risk and potential losses.

Possibilistic C-Means (PCM) algorithm has been a cornerstone in fuzzy clustering, particularly valued in data mining and pattern recognition. PCM's applicability in image analysis and knowledge discovery is well-documented, but its efficacy diminishes with the increasing scale and heterogeneity of data. This literature survey explores the advancements in PCM to address these challenges, focusing on high-order PCM (HOPCM), distributed computing approaches, and privacy-preserving techniques.

Problem Statement

Traditional methods of predicting share market values often fall short due to the complex and volatile nature of the market. This unpredictability can result in heavy financial losses for investors who rely on inaccurate predictions. Therefore, there is a pressing need for a more robust and accurate prediction model to safeguard investors' interests.

Proposed Solution: C-Means Clustering Algorithm

The core of my project is the implementation of the C-Means (CM) clustering algorithm. CM is a popular data clustering technique that groups similar data points into clusters. By applying this algorithm to share market data, we can identify patterns and trends that are not immediately obvious through conventional analysis.

Key Features of C-Means Clustering:

- **Data Segmentation:** CM segments vast amounts of market data into distinct clusters, helping to identify and understand various patterns and trends within the market.

DOI: 10.48175/IJARSCT-18608

- **Churn Data Clustering**: The algorithm clusters churn data, which refers to data indicating potential exits from the market, aiding in the prediction of market movements and investor behavior.
- **Risk Reduction**: By providing clearer insights into market trends, CM reduces the risk for clients, enabling them to make more informed investment decisions.

Implementation and Benefits

Using the CM algorithm, my project will analyze historical market data to cluster similar patterns together. Market experts can then use these clusters to predict future market behavior with greater accuracy. Here's how it benefits the clients:

- **Enhanced Accuracy**: By clustering data, CM improves the precision of market predictions, providing clients with more reliable investment advice.
- **Informed Decision-Making**: Clients receive actionable insights based on comprehensive data analysis, helping them make better-informed decisions about buying and selling shares.
- **Reduced Risk**: By identifying high-risk clusters and potential market downturns, CM helps in minimizing the chances of financial loss, offering a protective measure against market volatility.

Traditional Possibilistic C-Means (PCM)

PCM, as introduced by Krishnapuram and Keller (1993), modified the fuzzy c-means (FCM) by incorporating a possibilistic approach to handle noise and outliers better. However, PCM's inherent design caters to small, structured datasets, limiting its performance in big data scenarios characterized by volume, variety, and velocity.

High-Order Possibilistic C-Means (HOPCM)

To extend PCM's capabilities to big data, recent research proposes the High-Order PCM (HOPCM) algorithm, which optimizes the objective function within the tensor space. This method leverages tensor algebra to handle multi-dimensional data effectively, capturing the underlying structure of heterogeneous datasets more accurately than traditional matrix-based methods.

Advantages of HOPCM:

Enhanced Data Representation: By operating in tensor space, HOPCM can represent and process multi-way data efficiently, leading to improved clustering performance.

Scalability: HOPCM addresses the scalability issues of PCM, making it suitable for larger datasets by exploiting higher-order relationships within the data.

Distributed HOPCM Using MapReduce

To further tackle the scalability challenge, HOPCM can be distributed across a cluster of machines using the MapReduce programming model, as initially conceptualized by Dean and Ghemawat (2004). This distributed approach allows HOPCM to process vast amounts of heterogeneous data by parallelizing computations and minimizing communication overhead.

Key Features:

- Parallel Processing: The MapReduce implementation of HOPCM divides the clustering task into smaller, manageable subtasks, processed in parallel across a distributed system.
- Fault Tolerance: The MapReduce framework inherently supports fault tolerance, ensuring reliable processing even in the presence of hardware failures.

Privacy-Preserving HOPCM (PPHOPCM)

In the era of cloud computing, data privacy is paramount. The Privacy-Preserving HOPCM (PPHOPCM) algorithm integrates the BGV (Brakerski-Gentry-Vaikuntanathan) encryption scheme to secure the clustering process. The BGV scheme, a homomorphic encryption method, allows computations on encrypted data without revealing the actual data.

DOI: 10.48175/IJARSCT-18608

Implementation Details:

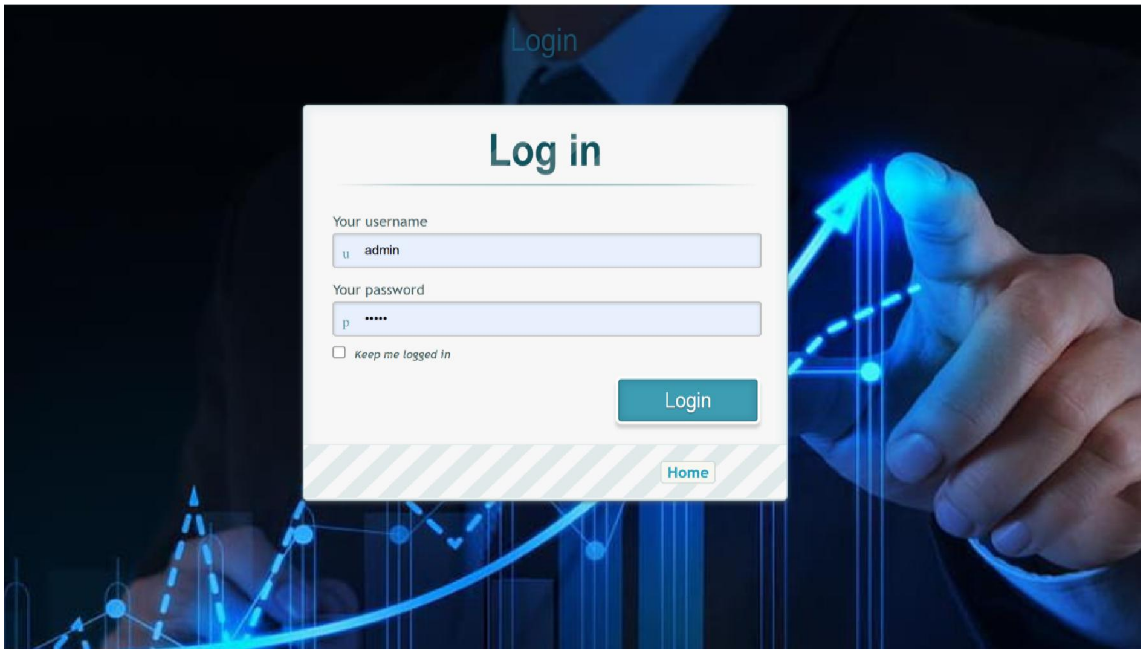
- Encrypted Computations: PPHOPCM approximates functions for updating the membership matrix and clustering centers as polynomial functions to support the BGV scheme's secure computing capabilities.
- Data Confidentiality: By encrypting data before processing, PPHOPCM ensures that sensitive information remains protected throughout the clustering process.
- Experimental Validation:
- Effectiveness: Experimental results demonstrate that PPHOPCM can successfully cluster large heterogeneous datasets without compromising data privacy.
- Efficiency: Despite the computational overhead of encryption, PPHOPCM maintains a practical balance between security and performance, making it viable for real-world applications.

III. CONCLUSION

By integrating the C-Means Clustering algorithm into market analysis, my project aims to significantly enhance the accuracy of share market predictions. This method provides a data-driven approach to understanding market trends, thereby reducing investment risks and safeguarding clients from potential losses. With this system, clients can invest with greater confidence, guided by the precise and reliable predictions of market experts.

In summary, this project not only addresses the limitations of traditional prediction models but also offers a sophisticated solution that leverages advanced clustering techniques to improve market predictions.a robust solution for investors looking to navigate the complexities of the share market.

Fig.1 Login page

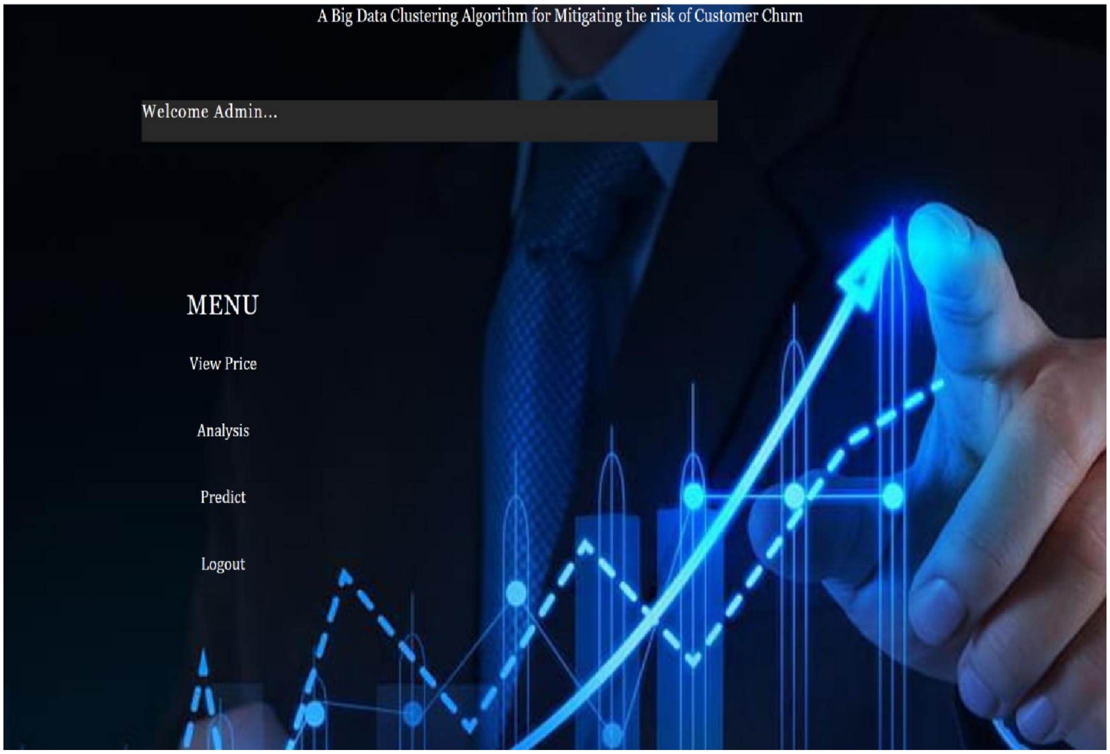


Your username
admin

Your password
•••••

DOI: 10.48175/IJARSCT-18608

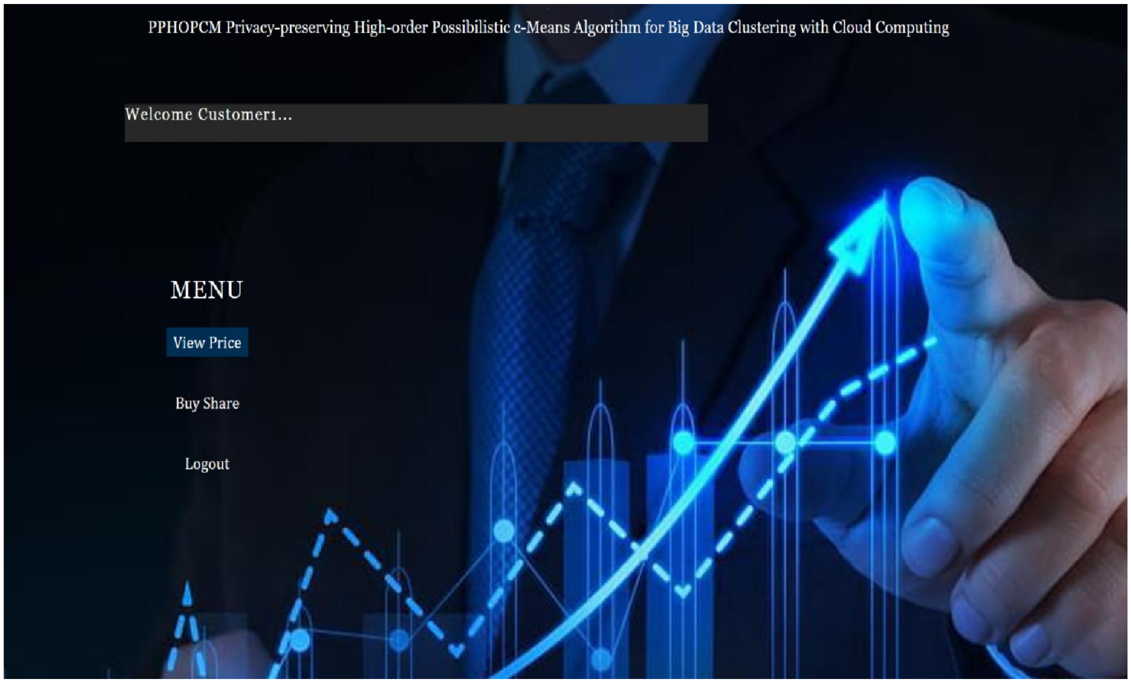
Fig.2 Admin page



A Big Data Clustering Algorithm for Mitigating the risk of Customer Churn

Welcome Admin...

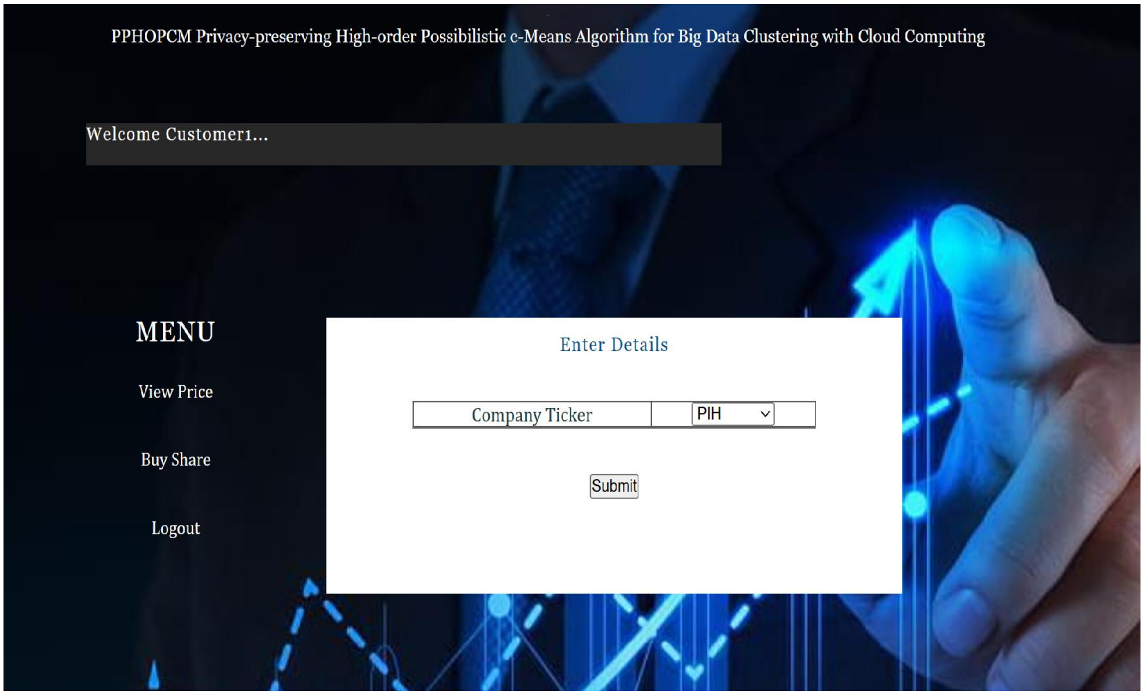
Fig.3 Menu page



PPHOPCM Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

Welcome Customer1...

Fig.4 View Price

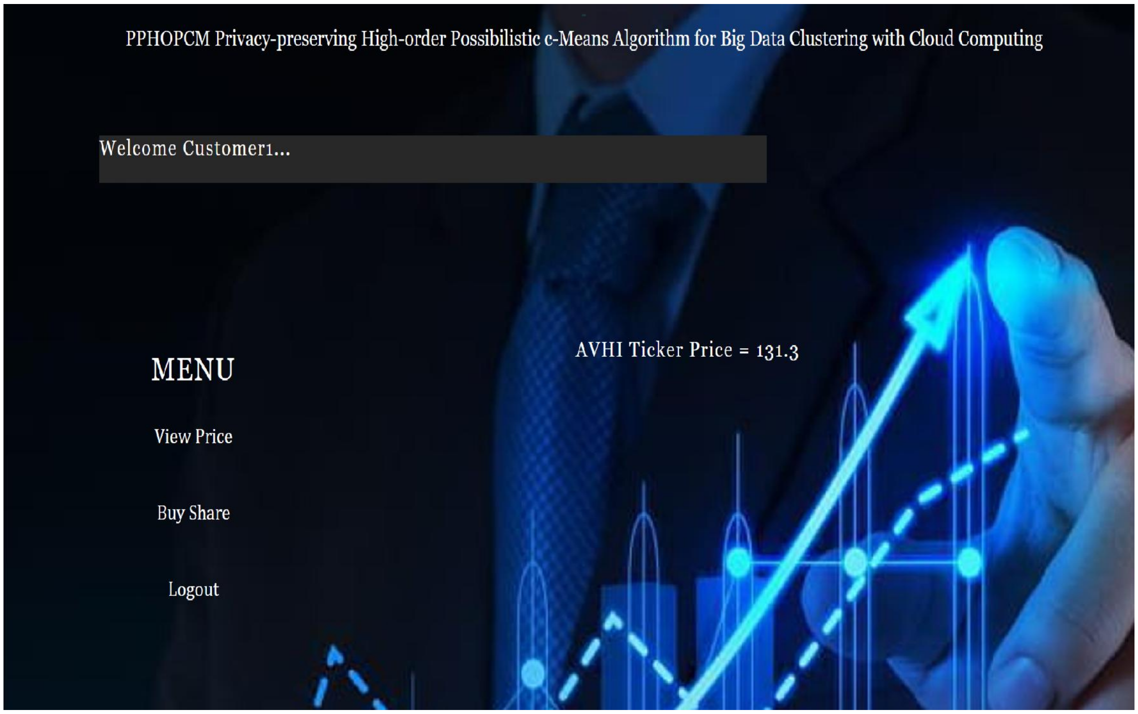


PPHOPCM Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

Welcome Customer1...

Company Ticker: PIH

Fig.5 Price Setting

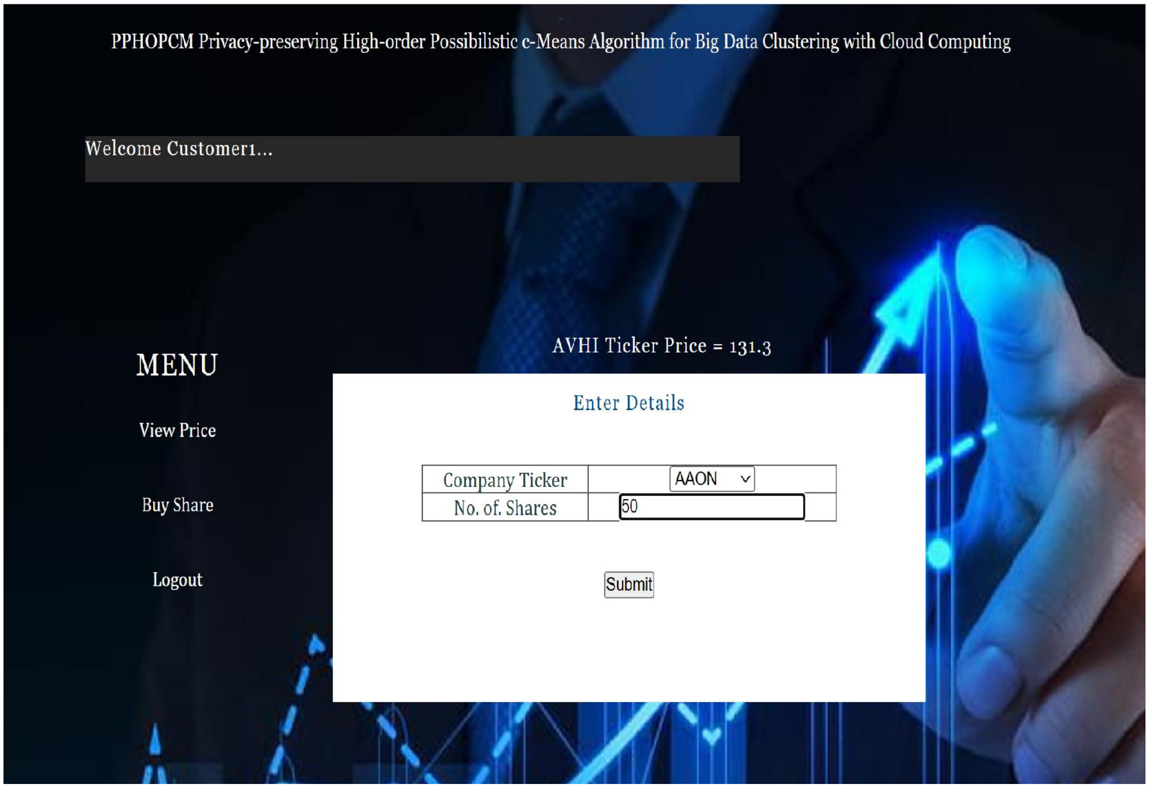


PPHOPCM Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

Welcome Customer1...

AVHI Ticker Price = 131.3

Fig.6 View Price Showing



PPHOPCM Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

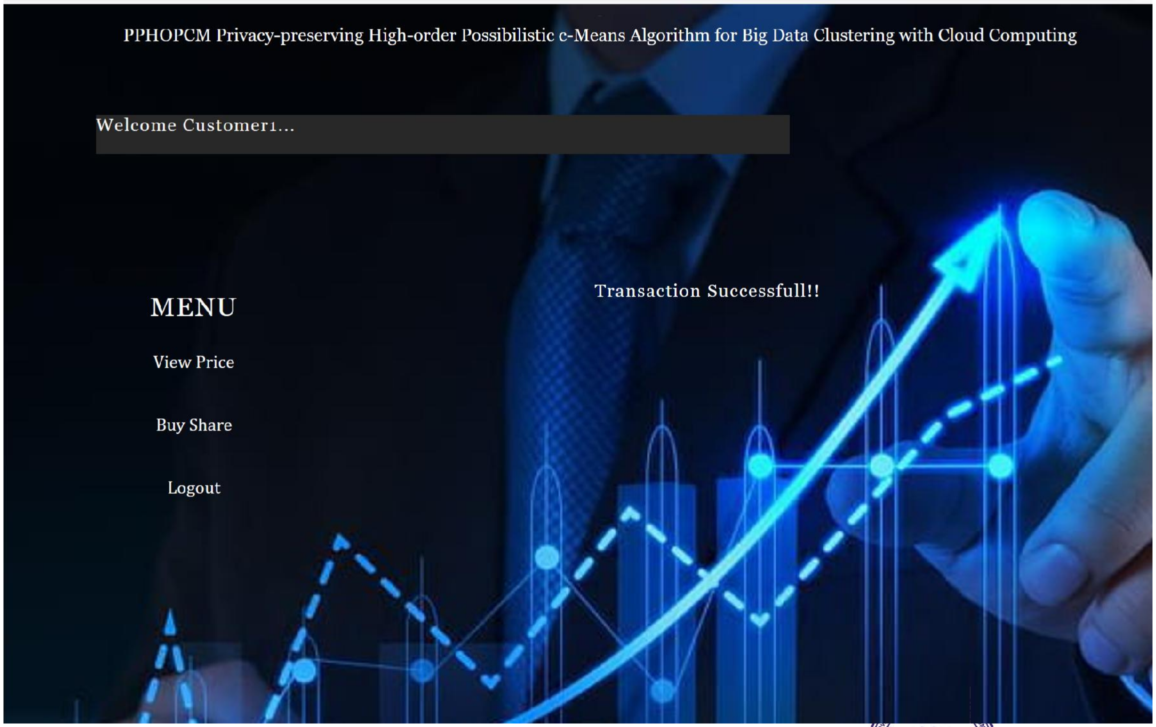
Welcome Customer1...

AVHI Ticker Price = 131.3

Enter Details

Company Ticker	AAON
No. of. Shares	50

Fig.7 Buy Share



PPHOPCM Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

Welcome Customer1...

Transaction Successfull!!

DOI: 10.48175/IJARSCT-18608

IV. CONCLUSION

We presented a high-order PCM approach for heterogeneous data clustering in this study. In addition, cloud servers are used to create a distributed HOPCM scheme based on MapReduce, which increases big data clustering efficiency. One feature of the paper is the development of a privacy-preserving HOPCM algorithm for cloud privacy using the BGV approach. Based on experimental results, PPHOPCM can securely cluster large amounts of data using cloud computing technology without compromising user privacy. In reality, because DHOPCM is more efficient than PPHOPCM, it is better appropriate for large-scale

Given that PPHOPCM and DHOPCM exhibit strong scalability, as is verified by the experimental results, their efficiency can be further enhanced by utilizing additional cloud servers, rendering them more appropriate for big data clustering. heterogeneous data that does not need to be protected.

REFERENCES

[1] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, “Computer assisted customerchurn management: State-of-the-art and future trends,” Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, Oct. 2007.

[2] N. Lu, H. Lin, J. Lu, and G. Zhang, “A customer churn prediction model in telecom industry using boosting,” IEEE Trans. Ind. Informat., vol. 10, no. 2, pp. 1659–1665, May 2014.

[3] B. Q. Huang, T. K. Mohand, and B. Brian, “Customer churn prediction in telecommunications,” Expert Syst. Appl., vol. 39, no. 1, pp. 1414–1425, Jan. 2012.

[4] E. G. Castro and M.S.G. Tsuzuki, “Churn prediction in online games using players’ login records: A frequency analysis approach,” IEEE Trans. Comput. Intell. AI Games, vol. 7, no. 3, pp. 255–265, Sep. 2015.

[5] W. H. Au, K. C. C. Chan, and Y. Xin, “A novel evolutionary data mining algorithm with applications to churn prediction,” IEEE Trans. Evol. Comput., vol. 7, no. 6, pp. 532–545, Dec. 2003.

[6] S. Y. Hung, D. C. Yen, and H. Y.Wang, “Applying data mining to telecom churn management,” Expert Syst. Appl., vol. 31, no. 3, pp. 515–524, Oct. 2006.

[7] T. Verbraken, V. Wouter, and B. Bart, “A novel profit maximizing metric for measuring classification performance of customer churn prediction models,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 5, pp. 961–973,May 2013.

[8] Y. Huang et al., “Telco churn prediction with big data,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, San Francisco, CA, USA, 2015, pp. 607–618.

[9] C. L. Chen and CY. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” Inf. Sci., vol. 275, pp. 314–347, Aug. 2014.

[10] H. Li, D. Wu, and G. X. Li, “Enhancing telco service quality with big data enabled churn analysis: Infrastructure, model, and deployment,” J. Comput. Sci. Technol., vol. 30, no. 6, pp. 1201–1214, Nov. 2015.

DOI: 10.48175/IJARSCT-18608