A Cluster Analysis of Investment Strategies In The Share Market

Akshaya .S[1], Rinu Stesha George[2], Jayashankari .J[3], Veeralakshmi .P[4]
Dept of Information Technology
Prince Shri Venkateshwara Padmavathy Engineering College , Chennai.

Abstract- *This paper presents big data system. In existing system, the new clustering algorithm called K-Means Clustering method (KMC) is used. It analysis the Churn data in Map Reduce function using the Possibilistic C-Means algorithm (PCM). The High-Order Possibilistic C-Means algorithm (HOPCM) for big data clustering used to optimizing the objective function on the tensor space. After clustering we analysis the Churn Data in Map Reduce Function. A Map-Reduce program consisting of a Mapper() that performs filtering and sorting (such as sorting tickets by first name into queues, one queue for each name) and a Reducer() that performs some operation (such as counting the number of tickets in each queue, yielding name frequencies). This project is used to find the profit and loss for the customers share based on clustering analysis technique for the particular tickers. In the existing system the large amount of data cannot be clustered at a time. So it divide and analyze the data to predict the profit and loss for the share. Then it combines the predicted results and produce the final result. It takes more time because the clustering and analysis process is done twice for the single ticker. The final result which was predicted was not accurate. To overcome this problem we proposed a new technique in which clustering and analysis for a large amount of data can be predicted at the single process. The result obtained was also accurate. It takes less time for clustering and analysis as this process occurs only once. And at last it clusters total no. of customers for that particular tickers and the shares invested by them and also the minimum and maximum share for the particular tickers in High-order Possibility (Maximum to minimum like descending order) by using c-Means Algorithm. The main disadvantage in the existing system is that we have to move to separate window to view the price for each tickers but in our proposed system we can able to view the price of each tickers by selecting them in the same window.*

Keywords- Big data clustering, Cloud computing, Privacy preserving, Possibilistic C-Means, Tensor space.

I. INTRODUCTION

AS personal computing technology and social websites, such as Facebook and Twitter, become increasingly popular, big data is in the explosive growth. Big data are typically heterogeneous,

i.e., each object in big data set is multi-modal Specially, big data sets include various interrelated kinds of objects, such as texts, images and audios, resulting in high heterogeneity in terms of structure form, involving structured data and unstructured data.

Moreover, different types of objects carry different information while they are interrelated with each other. 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.

Clustering is designed to separate objects into several different groups according to special metrics, making the objects with similar features in the same group. Clustering techniques have been successfully applied to knowledge discovery and data engineering. With the increasing popularity of big data, big data clustering is attracting much attention from data engineers and researchers. For example, Gao et al, designed a graph-based co-clustering algorithm for big data by generalizing their previous image-text clustering method. Chen et al,designed a nonnegative matrix tri-factorization algorithm to cluster big data sets by capturing the correlation over the multiple modalities.Zhang et al. Proposed a high-order clustering algorithm for big data by using the tensor vector space to model the correlations over the multiple modalities. However, it is difficult for them to cluster big data effectively, especially heterogeneous data, due to the following two reasons. First, they concatenate the features from different modalities linearly and ignore the complex correlations hidden in the heterogeneous data sets, so they are

not able to produce desired results. Second, they often have a high time complexity, making them only applicable to small data sets. Thus, they cannot cluster large amounts of heterogeneous data efficiently.

To tackle the above problems, this paper proposes a privacy preserving high-order PCM scheme (PPHOPCM) for big data clustering. PCM is one important scheme of fuzzy clustering. PCM reflects the typicality of each object to different clusters efficiently and also to avoid the corruption of noise in the clustering process.

II. RELATED WORK

2.1. A High-Order Possibilistic-Means Algorithm for Clustering Incomplete Multimedia Data

Clustering is a commonly used technique for multimedia organization, analysis, and retrieval. It is too difficult to capture the high-order nonlinear correlations over multimodal features, which results in lower accuracy clustering and they cannot extract features from the multimedia data with missing values, that leads to the failure in clustering incomplete multimedia data. This, proposes a HOPCM algorithm for clustering incomplete multimedia data. HOPCM improves the basic autoencoder model for learning features of multimedia data with missing set of values. So they use HOPCM tensor distance instead of the Euclidean distance as the exact distance metric for capturing the possibility of the unknown high-dimensional distribution of multimedia data. The results demonstrate that HOPCM achieves significantly better clustering performance than many existing algorithms. Most importantly, HOPCM is able to cluster both high-quality multimedia data and incomplete multimedia data efficiently, while the other existing methods can only cluster the high-quality multimedia data.

2.2. The Possibilistic c-Means Algorithm: Insights and Recommendations

The PCM algorithm is used to address the drawbacks associated with the constrained memberships in Fuzzy C-Means (FCM) algorithm. In this issue, Barni et al. (1996) report a difficulty they faced while applying the PCM, and note that it exhibits an undesirable tendency to converge to coincidental clusters. The purpose of this paper is not just to address the issues raised by Barni et al., but to go further and analytically examines the underlying principles of the PCM and the possibilistic approach, in general. We analyze the data sets used by Barni et al. and interpret the results reported by them in the light of our findings.

2.3. Applying the Possibilistic c-Means Algorithm in Kernel-Induced Spaces

The kernel extension of the classic possibilistic *c*-means. In the proposed extension, we implicitly map input patterns into a possibly high-dimensional space by means of positive semi definite kernels. The new space is for modeling the mapped data by means of the possibilistic clustering algorithm. We study in more detail the special case where we model the mapped data using a single cluster only, since it turns out to have many interesting properties. The modeled memberships in kernel-induced spaces yield in a modeling of generic shapes in the input space. We analyze in detail the connections to one-class support vector machines and kernel density estimation, thus, suggesting that the proposed algorithm can be used in many scenarios of unsupervised learning. In the experimental part, we analyze the stability and the accuracy of the proposed algorithm on some synthetic and real datasets. The results show high stability and good performances in terms of accuracy.

2.4 . A Tensor-based Approach for Big Data Representation and Dimensionality Reduction

It has been a great challenge to efficiently represent the process big data with a unified scheme. The unified tensor model is proposed to represent the unstructured, semi structured, and structured data. With tensor extension operator, various types of data are represented as sub tensors and then are merged to a unified tensor. In order to extract the core tensor which is small but contains valuable information, an incremental high order singular value decomposition (IHOSVD) method is presented. By recursively applying the incremental matrix decomposition algorithm, IHOSVD is able to update the orthogonal bases and compute the new core tensor. Analyzes in terms of time complexity, memory usage, and approximation accuracy of the proposed method are provided in this paper. A case study illustrates that approximate data reconstructed from the core set containing 18% elements can guarantee 93% accuracy in general. Analyzes and experiments results in demonstrating the proposed unified tensor model and IHOSVD which are more efficient for big data representation and dimensionality reduction.

2.5. Weighted Possibilistic c-Means Clustering Algorithms

This paper proposes the weighted possibilistic c-means algorithm. The weights indicate the possibility of a given feature vector belongs to any cluster. By assigning low weight values to outliers, the effects of noisy data on the clustering process is reduced. It is shown that the possibilistic

c-means algorithm is a special case of the weighted possibilistic c-means algorithm if each feature vector weight is assigned to one. Several methods for determining the weight values are presented. The performance of the algorithms is tested using data generated by a Gaussian random number generator with outliers and an artificial data set containing outliers.

2.6. MOiD (Multiple Objects incremental DBSCAN) – A paradigm shift in incremental DBSCAN

Mining an unprecedented increasing volume of data is a herculean task. Many mining techniques are available and being proposed every day. Clustering is one of those techniques used to group unlabeled data. Among prevailing proposed methods of clustering, DBSCAN is a density based clustering method widely used for spatial data. The major problems of DBSCAN algorithm are, its time complexity, handling of varied density datasets, parameter settings etc. Incremental version of DBSCAN has also been proposed to work in dynamic environment but the size of increment is restricted to one data object at a time. This paper presents a new flavour of incremental DBSCAN which works for multiple data objects at a time, named MOiD (Multiple Objects incremental DBSCAN). MOiD has been experimented on thirteen publicly available two dimensional and multi-dimensional datasets. The results show that MOiD performs significantly well in terms of clustering speed with a minor variation in accuracy.

III. POSSIBILISTIC K-MEANS ALGORITHM

The new clustering algorithm called K-Means clustering method (KMC) is used. It analysis the Churn data in Map Reduce function. At first, we cluster the churn data using K-means algorithm. Clustering is the process of partitioning a group of data points into a small number of clusters.

$$Jm(U,V) = c \sum i=1n \sum j=1um\ ij \|xj - vi\|2 + c \sum i=1 \eta in \sum j=1(1-uij)m, \quad (1)$$

where $V = \{v1; \cdots ; vc\}$ represents the set of clustering centers, *uij* denotes the membership of *xj* belonging to *vi*.

By minimizing Eq. (1), the membership matrix and the clustering centers can be updated by Eq.(2) and Eq.(3).

$$uij = 1\ (1+(d2\ ij/\eta i)1/(m-1)), \forall i,j, \quad (2)$$

$$vi = \sum n\ j=1\ um\ ijxj \sum n\ j=1\ um\ ij, \quad (3)$$

where *dij* denotes the distance between the *j*-th object *xj* and the *i*-th clustering center *vi*, and _*i* is a scale parameter which can be estimated by using Eq.(4):

$$\eta i = \sum n\ j=1\ um\ ij \times d2\ /ij \sum n\ j=1\ um\ ij \quad (4)$$

Typically, the computational complexity of the traditional possibilistic c-means algorithm is dominated by calculating the distance between each object *xj* and every clustering center *vi*, which requires $o(n \times c)$ for each iteration. So, PCM has a computational complexity of o(tn*c) with t indicating the number of iterations.

IV. BIG DATA CLUSTERING

Over the past few years, some algorithms have been proposed for big data clustering, especially for heterogeneous data sets. Early works focused on image-text co-clustering by information fusion. Specially, many algorithms first extracted the image features and the text features separately, and then concatenated them into a single vector. However, these methods are difficult to produce desired clustering results since they cannot capture the complex correlations over the bi-modalities of the objects by concatenating the features in linear way. To tackle this problem, Jiang and Tan . Proposed two methods based on the vague information and the Fusion ART to learn the visual-textual correlations by measuring the image-text similarities.

Most of heterogeneous data clustering schemes are developed depending on graph theory. They usually transform the heterogeneous data clustering task into a graph partitioning problem. The most representative scheme of this type is the bipartite graph partition scheme proposed by Gao for image-text clustering by interpreting the clustering task as a tripartite graph. Afterward, they extended this method for heterogeneous data clustering. The similar work is the isoperimetric co-clustering algorithm proposed by Rege et al. This algorithm clusters heterogeneous data by solving a set of linear equations. In addition, Cai et al developed a spectral clustering algorithm, which is a representative method based on graph theory, for heterogeneous data clustering by designing an iterative process to optimize a unified objective function. Another graph theory based method is spectral relational clustering (SRC) presented by Long et al. SRC first produces a collective clustering and then achieves the final result by deriving an iterative spectral clustering. The major weakness of the heterogeneous data clustering algorithms based on the graph theory is that they have a high time complexity. Therefore, they are often inefficient for large amounts of data.

Another kind of heterogeneous data clustering is based on the matrix factorization theory. For instance, Chen et al presented a clustering algorithm based on non-negative matrix factorization scheme for heterogeneous data clustering by minimizing the global reconstruction function of the relational matrix over multiple modalities. Other heterogeneous data clustering techniques, such as the combinatorial Markov random fields (Comrafs), are depending on information theory. Similar to the methods based on graph theory, these algorithms still have a high time complexity. For example, the computational complexity of Comrafs will increase significantly with the growing amount of heterogeneous data.

V. HIGH-ORDER POSSIBILISTIC C-MEANS ALGORITHM

The user will be able to take decision about buying stock shares by providing a exact information of feature share values in share market. In this project we created a database using stack details of several companies. So using this information the system can provide accurate share value of particular company . The algorithm used is possibilistic c-means algorithm(PCM).The high-order PCM algorithm (HOPCM) is proposed for big data clustering by optimizing the objective function in the tensor space**.**

$$Jm(U,V)=c\sum i=1n\sum j=1um\ ijd2\ (T)ij+c\sum i=1\eta in\sum j=1(1-uij)m, \quad (5)$$

where $m > 1$ denotes a fuzzification constant. Generally, $m \rightarrow 1$ results in approaching a hard cluster result while $m \rightarrow \infty$ causes a high level of fuzziness.

Eq. (5) shows that HOPCM has the similar objective function with PCM. However, different from PCM, $d2(T)ij$ indicates the distance between two tensors, namely the jth object xj and the ith clustering center vi.

To calculate the distance $d2(T)ij$ between xj and vi, we unfold each tensor $O \in RI1\times I2\times\cdots\times IT$ used to represent the object xjor the clustering center vi to its corresponding vector o. Specially, ol, the l-th item of the vector o, denotes $xji1i2:::iT$ Σ by $l = i1 + Tk=2\Pi k-1\ t=1\ It$. Thus, we can calculate the distance $d2(T)ij$ between xj and vi by an inner product as Eq. (6).

$$d2\ (T)ij=\sum I1\times I2\times\cdots IT\ l=1\ (xjl-vil)2. \quad (6)$$

The goal of the high-order possibilistic c-means algorithm is to minimize the objective function (5). To update the membership value uij , we differentiate (5) with respect to uij and we can get:

$$\partial Jm(U,V)\ \partial uij=\partial\ \partial uij\ (um\ ijd2\ (T)ij+\eta i(1-uij)m)=m\cdot d2\ (T)ij\cdot um-1\ ij+m\cdot\eta i(1-uij)m-1 \quad (7)$$

We can get the equation for updating uij by setting Eq. (7) to 0 as Eq. (8).

$$uij=1\ (1+(d2\ (T)ij/\eta i)1/(m-1)),\forall i,j \quad (8)$$

Similarly, we can get the equation for updating vi with the same format as Eq. (3).

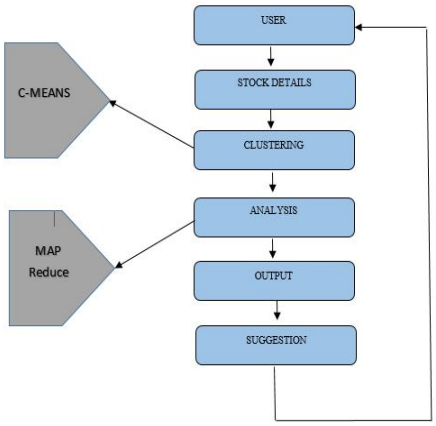

Fig 1 System Architecture

6.1 Authentication and Authorizations

In this module the User have to register first, then only he/she has to access the account. While registration the user can select the captcha and Image co-ordinates as they want. When the user login to the site, then he/she have to give correct password which they choose earlier. If the user gives wrong password, the account will be block. The authorization and authentication process facilitates the system to protect itself and besides it protects the whole mechanism from unauthorized usage. The Registration involves in getting the details of the users who wants to use this application.

6.2. Data Clustering Using C-Means

After Successful login, User can select the Company ticket to cluster the data. We use C-Means Cluster algorithm to cluster the data. Cluster analysis groups the data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and

different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering.

6.3.Analyzing using map-reduce

After Clustering, The next step is analysis using Map Reduce. Map-Reduce is a functional programming paradigm that is well suited to handling parallel processing of huge data sets distributed across a large number of computers, or in other words. Map-Reduce, as its name implies, works in two steps:

3.1.Map: The map step essentially solves a small problem: It divides the problem into small workable subsets and assigns those to map processes to solve.

3.2.Reduce: The reducer combines the results of the mapping processes and forms the output of the Map-Reduce operation.

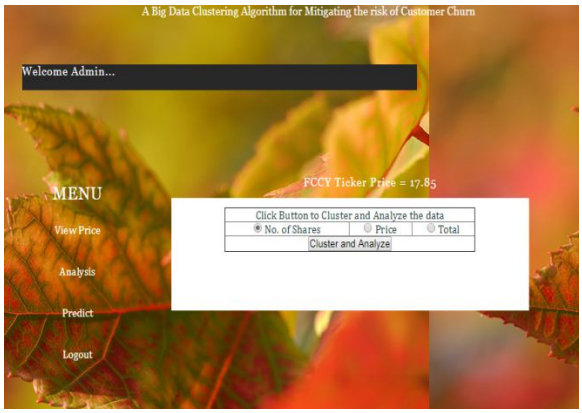

Fig 2 Map-reduce an analysis

```
Price: 9.5
Volume: 124648
P/E: 0.0
EPS: 0.65
Year Low: 7.8
Year High: 11.4
Day Low: 9.45
Day High: 9.7
50 Day Moving Av: 9.61
Market Cap: 0.0
The full name is: 1-800 FLOWERS.COM, Inc.
The currency is: USD
The short ratio is: 9.82
The previous close was: 9.6
The open for today was: 9.65
The exchange is NMS
aaaaaaaaaaa103
Price: 9.5
Volume: 124648
P/E: 0.0
EPS: 0.65
Year Low: 7.8
Year High: 11.4
Day Low: 9.45
Day High: 9.7
50 Day Moving Av: 9.61
Market Cap: 0.0
The full name is: 1-800 FLOWERS.COM, Inc.
```

Fig 3 output for c-means algorithm

VII. CONCLUSION

In this paper, we proposed a high-order PCM scheme for heterogeneous data clustering. Furthermore, cloud servers are employed to improve the efficiency for big data clustering by designing a distributed HOPCM scheme depending on Map Reduce.

REFERENCES

[1] Q. Zhang, L. T. Yang, Z. Chen, and Feng Xia, ”A High-Order Possibilistic-Means Algorithm for Clustering Incomplete Multimedia Data,”*IEEE Systems Journal, 2015, DOI: 10.1109/JSYST.2015.2423499.*

[2] R. Krishnapuram and J. M. Keller, ”The Possibilistic c-Means Algorithm:Insights and Recommendations,” *IEEE Transactions on Fuzzy Systems,*vol. 4, no. 3, pp.3

[3] M. Filippone, F. Masulli, and S. Rovette, ”Applying the Possibilistic c-Means Algorithm in Kernel-Induced Spaces,” *IEEE Transactions on Fuzzy Systems, vol. 18, no. 3, pp. 572-584, Jun. 2010*85-393, Aug. 1996.

[4] Q. Zhang and Z. Chen, ”A Weighted Kernel Possibilistic c-Means AlgorithmBased on Cloud Computing For Clustering Big Data,” *International Journal of Communica*

[5] N. Soni and A. Ganatra, ”MOiD (Multiple Objects Incremental DBSCAN)- A Paradigm Shift in Incremental DBSCAN,” *International Journal of Computer Science and Information Security, vol. 14, no. 4, pp. 316-346,*2016.