International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Privacy Preserving High Order Expectation Maximization Algorithm for Big Data Clustering with Redundancy Removal

R. Sureka[1], P. Shanmugapriya[2]

[1]ME Student, Department of CSE, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamil Nadu, India

[2]Assistant Professor, Department of CSE, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamil Nadu, India

ABSTRACT

Cloud computing has become increasingly prevalent, providing end-users with temporary access to scalable computational resources. At a conceptual level, cloud computing should be a good fit for technical computing users. A heterogeneous cloud, on the other hand, integrates components by many different vendors, either at different levels (a management tool from one vendor driving a hypervisor from another) or even at the same level (multiple different hypervisors, all driven by the same management tool).Nowadays, a large number of heterogeneous data, often referring to big data, is generating from big storage, which requires novel models and technologies to process, especially clustering based computing, for the further promotion the design and applications of big data analytics. However, the heterogeneous data is usually very complex, which is composed of structured data and unstructured data, such as picture, text, pdf and video. In other words, the heterogeneous data contain multimodal between which there are nonlinear relationships. In the existing work, proposed a high-order possibilistic c-means algorithm by extending the conventional possibilistic c-means algorithm from the vector space to the tensor space for multimedia heterogeneous data clustering. Furthermore, employed cloud computing to improve the clustering efficiency for massive heterogeneous data. To protect the private data during clustering on cloud, proposed a privacy-preserving expectation maximization algorithm by using the asymmetric encryption scheme to encrypt the original data. The existing BGV scheme does not support the division operations and exponential operations that are used in the membership matrix updating function of the high-order fuzzy c-means algorithm. To address this problem, use the asymmetric encryption scheme to approximate the membership matrix updating function to a polynomial function.

Keywords : Heterogeneous database, Big data, Cloud computing, Privacy Preserving, Clustering

I. INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing application software's are inadequate to deal with them. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem." Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on." Scientists, business executives, practitioners of medicine,

CSEIT1183567 | Received : 10 May 2018 | Accepted : 25 May 2018 | May-June-2018 [(3)5 : 293-300]

advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, finance, urban informatics, and business informatics. Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data. The work may require "massively parallel software running on tens, hundreds, or even thousands of servers". What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration. Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.The basic big data framework is shown in fig 1.

Fig 1: Big data deployment model

II. RELATED WORK

Zhang, Qingchen, et al [1] proposes a privacy preserving high-order PCM scheme (PPHOPCM) for big data clustering. PCM is one important scheme of fuzzy clustering. PCM can reflect the typicality of each object to different clusters effectively and it is able to avoid the corruption of noise in the clustering process. However, PCM cannot be applied to big data clustering directly since it is initially designed for the small structured dataset. Specially, it cannot capture the complex correlation over multiple modalities of the heterogeneous data object. The paper proposes a high-order PCM algorithm by extending the conventional PCM algorithm in the tensor space. Tensor is called a multidimensional array in mathematics and it is widely used to represent heterogenous data in big data analysis and mining. In this paper, the proposed HOPCM algorithm represents each object by using a tensor to reveal the correlation over multiple modalities of the heterogeneous data object. To increase the efficiency for clustering big data, we design a distributed HOPCM algorithm based on MapReduce to employ cloud servers to perform the HOPCM algorithm.

Q. Zhang,et.al,…[2] focus on the feature learning on big data. The characteristics of big data make the feature learning such an extremely challenging task that conventional data mining methods are almost impossible to address. On one hand, the unprecedented data volumes require the scalability of feature learning algorithms for big data. On the other hand, the high variety demands the feature learning algorithms which can learn the complex correlations among heterogeneous data to form an effective representation of big data. Therefore, it needs urgently new innovative theories and advanced technologies for feature learning on big data. Today, deep learning, together with advances in high performance computing, provides a new innovative solution for this problem. Deep learning refers to a set of machine learning models that perform supervised/unsupervised feature learning to automatically form hierarchical representations in deep architectures. The most typical deep learning model is the stacked auto-encoder (SAE) that is established by stacking auto-encoders such as Restricted Boltzmann Machines, sparse auto-encoders, denoising auto-encoders and predictive sparse coding.

L. Zhao, and P. Li, et.al,…[3] proposed a latest clustering algorithm based on fast finding and searching of density peaks (CFS) is improved to cluster a large number of dynamic data in industrial Internet of Things. It can find clusters of arbitrary shape and determine the number of clusters automatically. Some experiments have demonstrated its superiority in the efficiency and effectiveness over the previous algorithms for clustering large amounts of data. However, it is initially designed for only static data, making it limited for dynamic data clustering in industrial application. This paper aims to propose an incremental variant of CFS clustering, which can modify the current clustering results according to new arriving objects effectively and efficiently, rather than re-implement CFS clustering on the whole dataset. Therefore, a great deal of time can be saved, making CFS efficient enough to be used for industrial applications. Specially, an incremental CFS algorithm based on k-mediods (ICFSKM) is designed for clustering dynamic data collected from industrial Internet of Things.

X. Zhang, et.al,…[4] propose a new Bayesian Discriminative MTC (DMTC) framework. We implement two DMTC objectives by specifying the framework with four assumptions. The objectives are formulated as difficult Mixed Integer Programming (MIP) problems. We relaxed the MIP problems to two convex optimization problems. The first one, named convex Discriminative Multitask Feature Clustering (DMTFC), can be seen as a technical combination of the convex supervised Multitask Feature Learning (MTFL) and the Support Vector Regression based Multiclass MMC (SVR-M3C). The second one, named convex Discriminative Multitask Relationship Clustering (DMTRC), can be seen as a technical combination of the convex Multitask Relationship Learning (MTRL) and SVR-M3C. These combinations are quite natural and yield the following advantages: In respect of "what to learn", DMTFC can learn a shared feature representation between tasks. DMTRC can minimize the model differences of the related tasks. Both of them work in Frobenius norms under the regularization framework. In respect of "when to learn", DMTRC can learn the task relationship automatically from the data by learning the inter-task covariance matrix. In respect of "how to learn", both algorithms are generated from the Bayesian framework.

Y. Chen, et.al,…[5] propose a Semi-supervised NMF (SS-NMF) based framework to incorporate prior knowledge into heterogeneous data co-clustering. In the proposed SS-NMF co-clustering methodology, users are able to provide constraints on data samples in the central type, specifying whether they "must" (must-link) or "cannot" (cannot-link) be clustered together. Our goal is to improve the quality of co-clustering by learning a new distance metric based on these constraints. Using an iterative algorithm, we then perform tri-factorizations of the new data matrices, obtained with the learned distance metric, to infer the central data clusters while simultaneously deriving the clusters of related feature modalities. The preliminary version of this work was first presented in a shortened form as conference abstracts. We propose a novel algorithm for heterogeneous data co-clustering based on NMF. Computationally, NMF co-clustering is more efficient and flexible than graph-based models and can provide more intuitive clustering results. To the best of the knowledge, this is the first work on the semi-supervised co-clustering of multiple data types

III. EXISTING METHODOLOGIES

Clustering is an essential data mining and tool for analyzing big data. There are difficulties for applying clustering techniques to big data duo to new challenges that are raised with big data. As Big

Data is referring to terabytes and petabytes of data and clustering algorithms are come with high computational costs. These big data are still unloadable in the modern day computers. There are many approaches applied for clustering the big data. One technique is distributed clustering, where the data is distributed to various systems and the clustering is done independently on each system. K-means (or alternatively Hard C-means after introduction of soft Fuzzy C-means clustering) is a well-known clustering algorithm that partitions a given dataset into c (or k) clusters. It needs a parameter c representing the number of clusters which should be known or determined as a fixed apriori value before going to cluster analysis. Possibilistic C means clustering algorithm is a soft algorithm clustering fuzzy data in which an object is not only a member of a cluster but member of many clusters in varying degree of membership as well. In this way, objects located on boundaries of clusters are not forced to fully belong to a certain cluster, but rather they can be member of many clusters with a partial membership degree between 0 and 1. In spite of it's relatively higher cost in implementation. The encryption scheme can be implemented using BGV scheme which is a fully homomorphic encryption (FHE) that dramatically improves performance and bases security on weaker assumptions. A central conceptual contribution in work is a new way of constructing leveled fully homomorphic encryption schemes (capable of evaluating arbitrary polynomial-size circuits), without Gentry's bootstrapping procedure but provide large size of data

3.1 K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to recalculate k new centroids as barycentre's of the clusters resulting from the previous step.The improved K-means algorithm is a solution to handle large scale data, which can select initial clustering center purposefully, reduce the sensitivity to isolated point, and avoid dissevering big cluster. By using this technique locating the initial seed point is easy and which will give more accurate and high-resolution result. By using various techniques we can study or compare the results and find out which technique gives higher resolution Initial centroid algorithm is useful to avoid the formation of empty clusters, as the centroid values are taken with respect to the intensity value of the image. The basic algorithm pseudo code as follows:

Input: X = {x1, x2, x3,…..,xn} be the set of data points , Y= {y1,y2,y3…yn} be the set of data points and V = {v1,v2,v3,….,vn} be the set of centers

Step 1: Select 'c' cluster centers arbitrarily

Step 2: Calculate the distance between each data and cluster centers using the Euclidean Distance metric as follows

$$Dist(X,Y) = \sqrt{\sum_{j=1}^{n}(X_{ij} - Y_{ij})^2}$$

X, Y are the set of data points

Step 3: Pixel is assigned to the cluster center whose distance from the cluster center is minimum of all cluster centers

Step 4: New cluster center is calculated using

$$V_i = \frac{1}{C_i}\sum_{1}^{ci} x_i$$

Where Vi denotes the cluster center, ci denotes the number of pixels in the cluster

Step 5: The distance among every pixel and new obtained cluster facilities is recalculated

Step 6: If no data were reassigned then stop. Otherwise repeat steps from 3 to 5

3.2 Fuzzy K-means Clustering

Fuzzy K-means Clustering is a clustering algorithm in which each data point belongs to cluster to a degree specified by a membership grade. In this, object is grouped into K fuzzy groups. Cluster center is calculated for each group and the Euclidean distance is measured between the pixel and each centroid of clusters. Then the pixel is grouped with the cluster which has shortest distance to the centroid. FKM is a method of clustering which allows one pixel to belong to two or more clusters. The FKM algorithm attempts to partition a finite collection of pixels into a collection of K fuzzy clusters with respect to some given criterion. Depending on the data and the application, different types of similarity measures may be used to identify classes. Some examples of values that can be used as similarity measures include distance, connectivity and intensity.

The n sample of the data input data points is expressed as X = $\{x_1, x_2, \ldots, x_n\}$ while the corresponding cluster centres of the data points is expressed as V = $\{v_1 v_2, \ldots, v_c\}$, where c is the number of clusters. μ_{ij} is the membership degree of the image data point x_i to the cluster centre v_j Fuzzy clustering computes the optimum partition based on the minimization of the objective function given that μ_{ij} satisfies

$$\sum_{i=1}^{n} \mu_{ij} = 1, 1 \leq j \leq n$$

The cluster center (i.e centroid)V_j is computed as

$$V_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}}$$

Where m is the fuzziness index parameter and m $\in [1, \infty]$

Given that

$$d_{ij} = \|x_i - v_j\|$$

The dissimilarity between the centroids v_j and the data x_i is computed as

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m d_{ij}$$

Such that d_{ij} is the Euclidean distance between the i^{th}data point and the j^{th}centroid while $\mu_{ij} \epsilon [0,1]$ and the fuzziness index parameter m $\epsilon [1, \epsilon]$

The new membership value is further computed as

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} [\frac{d_{ij}}{d_{ik}}]^{\frac{2}{m-1}}}$$

this is iteratively computed until

$\left\| \mu_{ij}^{(k+1)} - \mu_{ij}^{(k)} \right\| < \lambda$

Where k is the iteration step and $\lambda \in [0,1]$ is the criterion for terminating the iteration

IV. PROPOSED METHODOLOGIES

Existing data mining techniques, more particularly iterative learning algorithms, become overwhelmed with big data. Big Data is has taken center stage in analytics. While there is no formal definition, and as many shades of interpretation as can be made from "data" and "big", there are several themes. Data clustering is a common computing task that often involves large data sets for which MapReduce can be an attractive means to a solution. In this project we can implement EM algorithm with Map Reduce framework in secure environments using El-gammal algorithm. Expectation-Maximization (EM) algorithm is an iterative technique for estimating the value of some unknown quantity, given the values of some correlated, known quantity. EM is generally preferable to K-means due to its better convergence properties to handle heterogeneous datasets. And also cluster the video using frames clustering approach. Then remove outliers using data deduplication approach based on variable chunk similarity algorithm. Data de duplication techniques ensure that only one unique instance of data is retained on storage media. Block-level deduplication looks within a file and saves unique iterations of each block. All the blocks are broken into chunks with the same fixed length. Instead of protecting the big data itself, the proposed scheme protects the mapping of the various data elements to each provider using El-Gammal algorithm. Analysis, comparison and simulation prove that the proposed scheme is efficient and secure for the big data of cloud tenants. Secured data transmission using ElGamal can be defined as transmission of data. El-gammal algorithm is a public key encryption technique based on elliptic curve theory that can be used to create faster, smaller, and more efficient cryptographic keys.

4.1 Expectation Maximization algorithm:

Expectation–maximization (EM) algorithm is an iterative method to find maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, the parameters and the latent variables, and simultaneously solving the resulting equations. In statistical models with latent variables, this is usually impossible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that there is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points. It's not obvious that this will work, but it can be proven that in this context it does, and that the derivative of the likelihood is (arbitrarily close to) zero at that point, which in turn means that the point is either a maximum or a saddle point.[12] In general, multiple maxima may occur, with no guarantee that the global maximum will be found. Some likelihoods also have singularities in them, i.e., nonsensical maxima. For example, one of the solutions that may be found by EM in a mixture model involves setting one of the components to have zero variance and the mean parameter for the same component to be equal to one of the data points.

4.2 El-gammal Algorithm:

El-gammal is a kind of public key cryptosystem like RSA. But it differs from RSA in its quicker evolving capacity and by providing attractive and alternative way to researchers of cryptographic algorithm. The security level which is given by RSA, can be provided even by smaller keys of El-gammal algorithm. For example, the 1024 bit security strength of a RSA could be offered by 163 bit security strength of El-gammal algorithm. Other than this, El-gammal algorithm is particularly well suited for wireless communications, like mobile phones, PDAs, smart cards and sensor networks. El-gammal point of multiplication operation is found to be computationally more efficient than RSA exponentiation. The El-gammal certainly is not the ellipse shape; they are so named because they are described by cubic equations, similar to those used for calculating the circumferences of points. The El-gammal algorithm can be shown in fig 2.

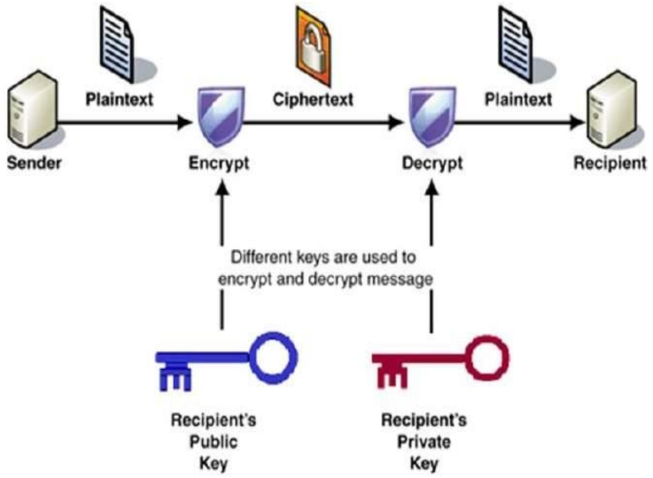

Fig 2: El-gammal Algorithm

The ElGamal algorithm can be use as RSA algorithm for public key encryption because:

RSA encryption depends on the difficulty of factoring large integers while ElGamal encryption depends on on the difficulty of computing dicrete logs in a large prime modulus. ElGamal is nothing but the advance version of Diffie-Hellmen key exchange protocol. But,ElGamalis not good because its cipher text is two times longer than the plain text. ElGamal is good because it gives different cipher text for same plain text each time.For image data, the size of the cipher text is very huge & reshaping the encrypted data was not understood. ElGamal's encryption is very simple because it is multiplication of message and symmetric key(i.e c=m*k). The proposed algorithm is shown in fig 3.

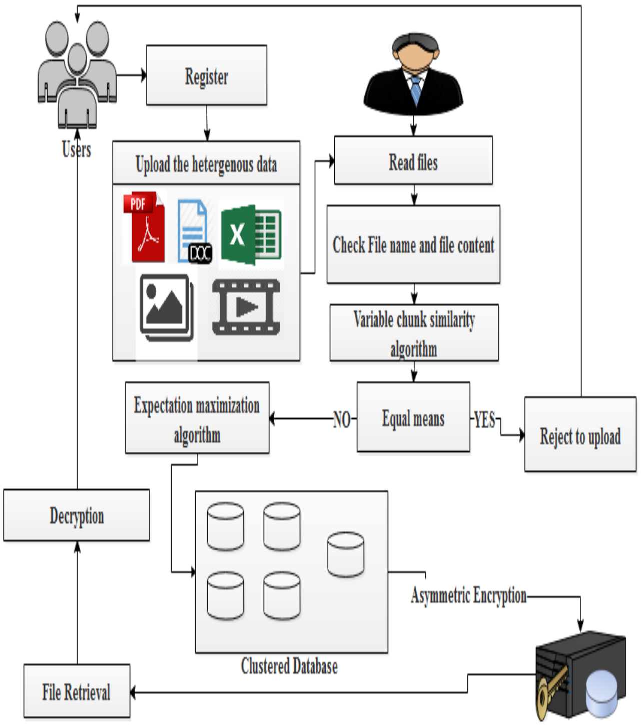
Users
Register
Upload the hetergenous data
Read files
Check File name and file content
Variable chunk similarity algorithm
Expectation maximization algorithm
NO
Equal means
YES
Reject to upload
Decryption
Clustered Database
Asymmetric Encryption
File Retrieval

Fig 3. Proposed Work

V. CONCLUSION

Big Data are the large amount of data being processed by the Data Mining environment. In other words, it is the collection of large and complex data sets which are difficult to process using traditional data processing applications. The clustering techniques are very useful to process data mining. Clustering is the process of grouping the data based on their similar properties. MapReduce is a feasible solution to processing problems involving large amounts of data. Especially for problems that can easily be partitioned into independent sub tasks that can be solved in parallel. We have presented an extension to the traditional clustering algorithm called EM that drastically reduced time complexity of PCM algorithm over large real world and synthetic data sets. EM presents an effective strategy to deal with scaling data and have begun employing this technique on different learning algorithms. In this work, we have identified a new privacy challenge during data accessing in the cloud computing to achieve privacy-preserving access authority sharing for similarity files by using variable chunk similarity algorithm. Authentication is established to guarantee data confidentiality and data integrity. Data anonymity is achieved since the wrapped values are exchanged during transmission with El-gammal algorithm. All the results show that the proposed scheme is effective and feasible to protect the big data for cloud tenants.

VI. REFERENCES

1. Chen. Y, L. Wang, and M. Dong, “Non-Negative Matrix Factorization for Semi supervised Heterogeneous Data Coclustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459-1474, Oct. 2010.
2. Jiang .T and A.-H. Tan, “Learning Image-Text Associations,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 2, pp. 161-177, Feb. 2009.
3. Long . B, X.Wu, Z. Zhang, and P. Yu, “Spectral Clustering for Multi-Type Relational Data,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 585-592.
4. Meng . L, A. Tan, and D. Xu, ”Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293-2306, Aug. 2014.
5. Zhang, Qingchen, et al. "PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing." IEEE Transactions on Big Data (2017).
6. Zhang. Q, L. T. Yang, and Z. Chen, “Deep Computation Model for Unsupervised Feature Learning on Big Data,” IEEE Transactions onServices Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016.
7. Zhang. Q, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, “An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things,” IEEE Transactions on Industrial Informatics, 2015.
8. Zhang Q, L. T. Yang, and Z. Chen, “Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning,” IEEE Transactions on Computers, vol. 65, no. 5, pp. 1351-1362, May 2016.
9. Zhao. R and W. Grosky, “Narrowing the Semantic Gap Improved Text- Based Web Document Retrieval Using Visual Features,” IEEE Transactions on Multimedia, vol. 4, no. 2, pp. 189-200, Jun. 2002.
10. Zhang . Q, L. T. Yang, Z. Chen, and Feng Xia, “A High-Order Possibilistic-Means Algorithm for Clustering Incomplete Multimedia Data,” IEEE Systems Journal, 2015