Elysium Journal of Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

A Weighted Kernel Possibilistic c-Means Algorithm Based On Cloud Computing For Clustering Big Data

M.Manjula

[1]*PG Student, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, TamilNadu, India*

Abstract—Here we will discuss about the Privacy-preserving High-order Possibilistic c-Means Algorithm. Fuzzy C- Means is a Clustering method that allows each data point to belong to multiple clusters with varying degree is membership. PCM is one of the methods used for C-means Clustering process and image analysis. The Process find out the two types of Clustering's like normal PCM clustering and important is HOPCM like (High Order PCM) FOR Big data clustering. The HOPCM method based on Map Reduce for very large amounts of heterogeneous data. Finally, a privacy-preserving HOPCM algorithm (PPHOPCM) to protect the private data on cloud by applying the BGV encryption scheme to HOPCM. To tackle this problem, the paper proposes a high-order PCM algorithm (HOPCM) for big data clustering by optimizing the objective function in the tensor space. 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.

Keywords—

1. INTRODUCTION

AS personal computing technology and social websites, such as Face book 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. Clustering is designed to separate objects into several different groups according to special metrics, making the objects with similar features in the same group. 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. 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 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 [14]. The paper proposes a high-order PCM algorithm by extending the conventional PCM algorithm in the tensor space. 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 Map Reduce to employ cloud servers to perform the HOPCM algorithm. However, the private data tends to be in disclosure when performing HOPCM on cloud. Take the medical data which is a typical type of big data for example. A large amount of private information such as personal email.

2. LITERATURE SURVEY

1.Zhenping Xie – Shitong Wang - F.L.Chung [5] in 2008 proposed a "An Enhanced Possibilistic c-Means clustering Algorithm EPCM" As an important data mining tool, the possibilistic c-means clustering algorithm (PCM) has emerged as an important technique for pattern recognition and data analysis proposed. Many PCM variants have been proposed alter standard PCM for improving the performance of the original PCM algorithm.2. Adam Schneider [19] in 2000 proposed a "Weighted possibilistic c-Means Clustering Algorithm" This paper proposed the weighted possibilistic c-means algorithm. It is difficulties the Fuzzy c-means algorithm has with noisy data. Further, this paper proposed the substitution of a set of membership weights into PCM objective function. The WPCM algorithm best for large set due to noise point is minimum. 3. Beyza Ermis, A. Taylan Cemgil, Evrim Acar.[2] in 2015 "Link Prediction in Heterogeneous Data Via Generalized Coupled Tensor" Link prediction is addressed coupled analysis of relational datasets including symmetric ones and multiway arrays Then they proposed tensor factorisation models i.e.., Generalised Coupled Tensor Factorisation (GCTF) .4. Y.Chen, L.Wang and M.Dong.[9] in 2010 "Non-Negative Matrix Factorization for Semi supervised Heterogeneous Data Coclustering" Coclustering heterogeneous data has attracted extensive various application are text mining, image retrieval, and Bioinformatics. In this paper proposed by Semi supervised Non-negative Matrix Factorization (SS-NMF) framework for data Coclustering. 5. X. Zhang[7] in 2005 "Convex Discriminative Multitask Clustering" Multitask clustering improve the clustering performance of multiple tasks. . In this

P-ISSN: 2347-4408
E-ISSN: 2347-4734

paper, we propose two convex Discriminative Multitask Clustering (DMTC) objectives to address the problems The first one aims combination of the convex multitask feature learning and the convex Multiclass Maximum Margin Clustering (M3C). The second one aims combination of the convex multitask relationship learning and M3C. The objectives of the two algorithms Bayesian framework. Finally Experimental results on a toy problem and two benchmark data sets. 6. X. Cai, F.Nie, H.Huang and F.kamanagar [25] in 1977-1984 "Heterogeneous image feature integration via multi-modal spectral clustering" In recent years, they proposed to describe objects and scenes appearing in images. In this paper, propose a novel approach to unsupervised integrate.

3. RELATED WORK

The research on possibilistic c-means (PCM) was first proposed by Krishnapuram and Keller in 1993. To overcome the weakness of Fuzzy c-means (FCM) they proposed in the year (1993, 1996). Since that time, the volume of research has grown tremendously: however, only in recent year the question about the experimental practices of privacy preserving high-order PCM (PPHOPCM) research come up.

Xie et al [2] developed an enhanced PCM algorithm by grouping the data set into one main subset and assistant subset to avoid the coincident result. In addition PCM is not robust to the addition parameter.

Yang et al [3] proposed an unsupervised PCM algorithm. To cluster non-spherical data sets, some kernel-based possibilistic clustering algorithms have been proposed by mapping the object of the data set into high order data space. Other PCM variants include weighted PCM algorithm [19] and sample-weighted PFCM algorithm.

The experiments on the two representative big data sets, i.e…., NUS-WIDE and SNAE2, to access the clustering accuracy and efficiency of our algorithm by comparison with three representative possibilistic c-means algorithms namely HOPCM-15, WPCM [3], PCM [1].

In between these surveys, several studies conducted more thorough analysis of the practise to overcome the weakness of the original PCM algorithm. Timm et al 2002[4] proposed two possibilistic fuzzy clustering algorithm that can avoid the coincident clustering problem of PCM by adding an inverse function of the distances between cluster center in PCM objective function.

Finally, we evaluate the scalability of DHOPCM and PPHOPCM in terms of speedup by performing DHOPCM and PPHOPCM in different in different platform 1 computer, 5 computers, 10 computers and 20 computers respectively. PPHOPCM can effectively cluster large number of heterogeneous data cloud computing without "disclosure" of private data.

4. EVALUATION METHODOLOGY

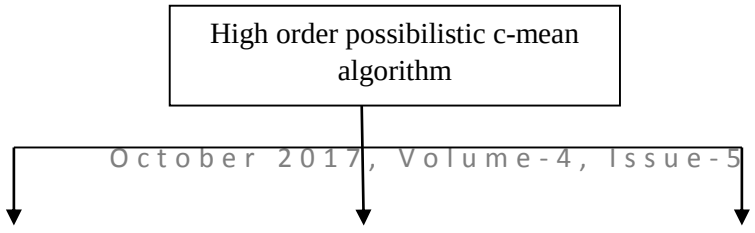
We conducted a survey of research work in the area of fuzzy possibilistic c-means algorithm published during the period of 1993; modified fuzzy possibilistic c means algorithm published during in the year 2003.To avoid the corruption of noise in the clustering. We collect all the research paper in Google Scholar and the Digital Bibliography and library project (DBLP) database for the reviewed time period. For this set, we excluded short paper, extended abstract and paper are not available in the English language.

4.1. Possibilistic C-Means (Pcm) Algorithm:

The possibilistic c-means (PCM) algorithm was proposed by Krishnapuram and Keller by removing the probabilistic constraint associated with the FCM algorithm [19]. Possibilistic c-means (PCM) relax the column sum constraint of fuzzy membership matrix in FCM and introduces a possibilistic partition matrix, so that possibilistic membership may reflect the typicalities of data points to their clusters well [5].

4.2. High Order Possibilistic C-Means Algorithm:

In this paper HOPCM algorithm based on future learning for clustering incomplete multimedia data. HOPCM implements three steps: unsupervised feature learning, feature fusion and high order clustering. Unsupervised feature learning: implement several feature learning/deep learning algorithm; feature fusion: The process of combining two or more distinct entities; high order clustering: high order clustering is also termed operators or functions.



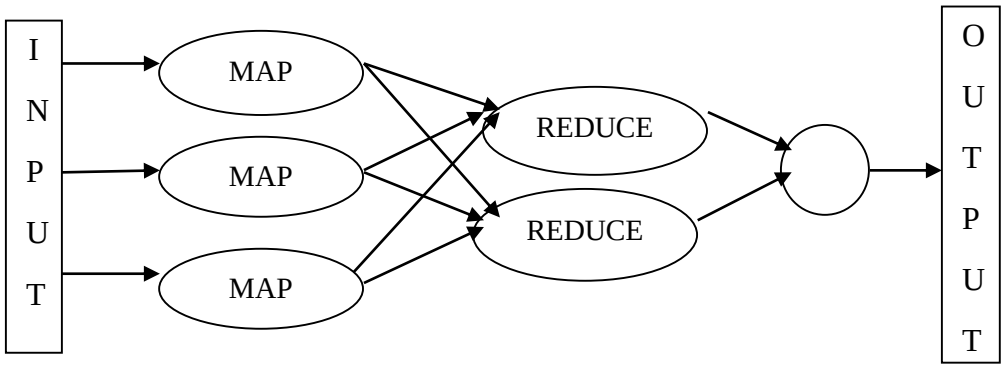
P-ISSN: 2347-4408
E-ISSN: 2347-4734

Unsupervised feature learning	Feature fusion	High order possibilistic c -means algorithm

Finally, a HOPCM algorithm is implemented for clustering the multimedia data in tensor space Future, we design distributed HOPCM based on Map Reduce for large amount of heterogeneous data.

4.3. Hadoop-Map Reduce:

Map Reduce is a processing technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce task is always performed after the map job.



Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

- Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
- Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

4.4. Privacy Preserving High Order Possibilistic C-Means Algorithm:

Finally, we devise a privacy preserving HOPCM algorithm (PPHOPCM) to protect the private data on cloud by applying the BGV encryption scheme to HOPCM.

Advantages:

- Proper secure for entire process.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

- The performance of communication is improved.
- Simple to store.
- Maintaining Feasibility.

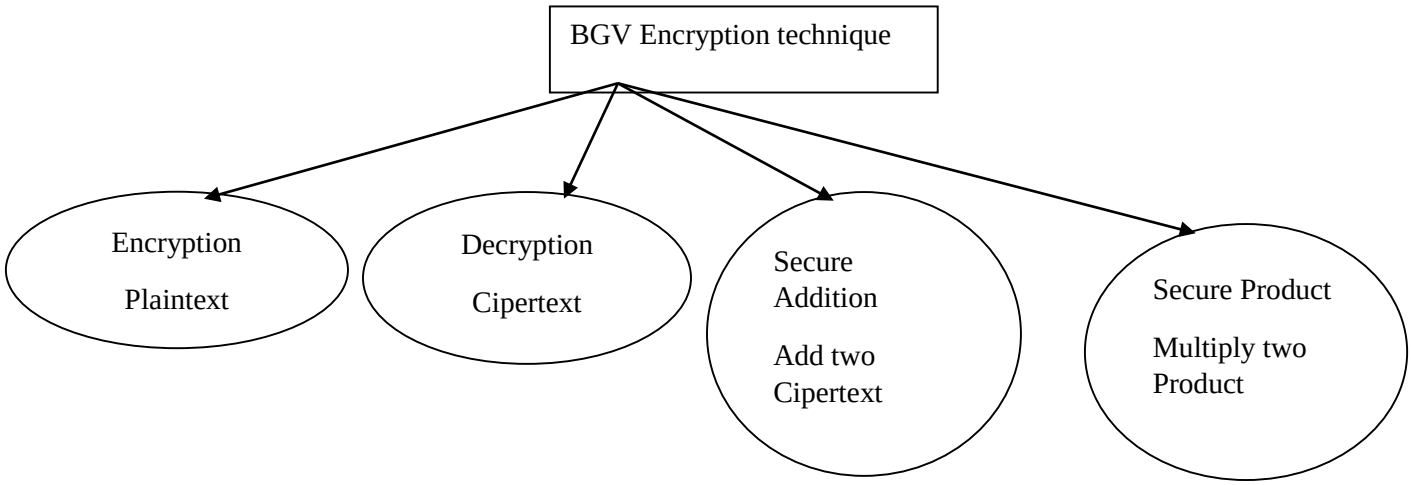
4.5. BGV encryption schema:

BGV is fully homomorphic encryption schema. Homomorphic encryption schemes are malleable by design. This enables their use in cloud computing environment for ensuring the confidentiality of processed data. In addition, the homomorphic property of various cryptosystems can be used to create many other secure systems, for example secure voting systems, collision-resistant hash functions, private information retrieval schemes.

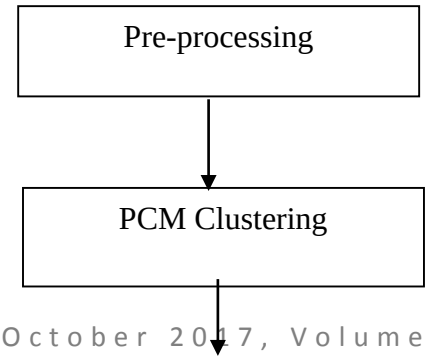
The BGV technique has four major operations:

Encryption

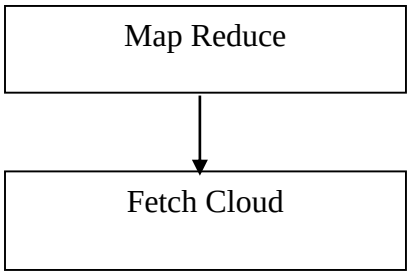
- Decryption
- Secure Addition
- Secure Product



5. PROPOSED SYSTEM

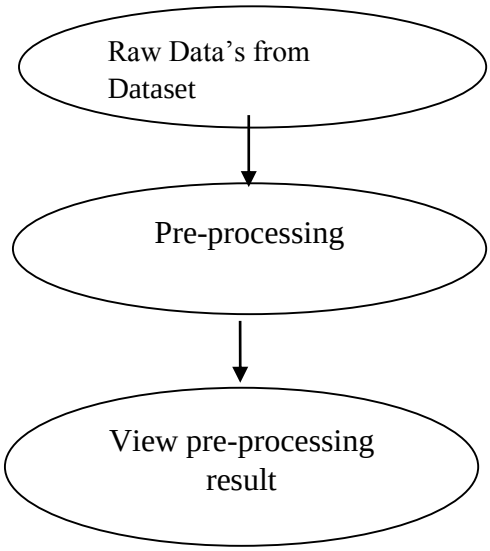


P-ISSN: 2347-4408
E-ISSN: 2347-4734



5.1. Preprocessing:

Pre-processing is one of main modules for data mining system. Here we are removing unwanted data or null values and unstructured data. So when we remove unstructured data's then only we get accurate results for given dataset. It is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values. Impossible data combinations, missing values, etc. Analysing data that has not been carefully screened for such problems can produce misleading results.

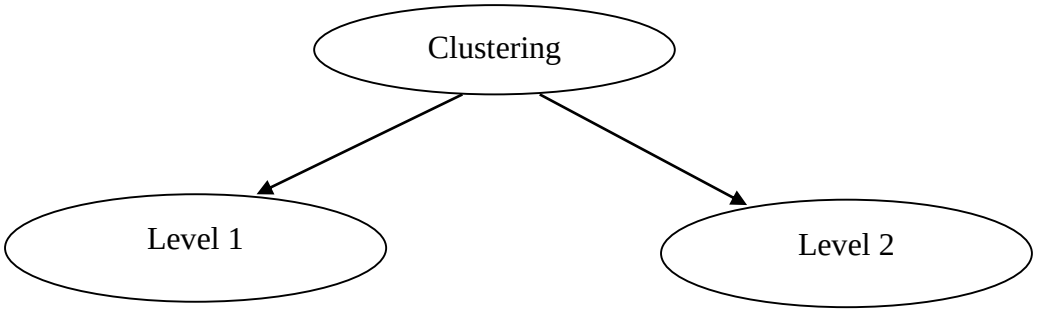


5.2. Pcm Clustering:

Here clustering is splitting data as some particular attributes based or analyzing attribute values through we will splitting or partitioning data individually. A fast PCM clustering algorithm is proposed in this

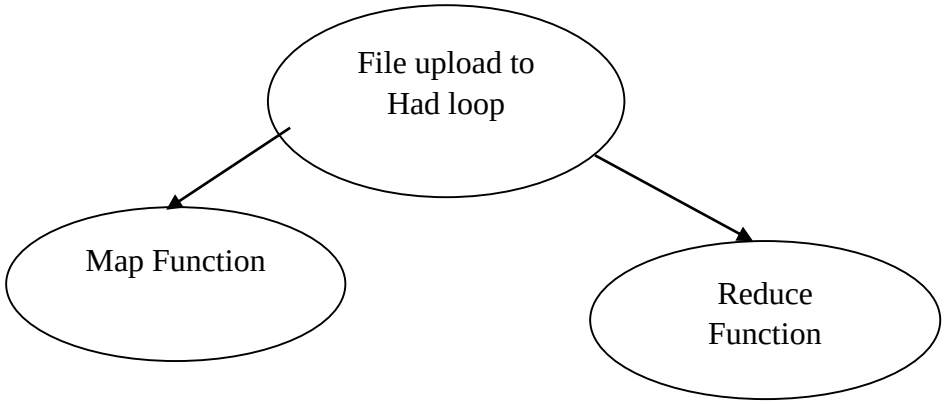
P-ISSN: 2347-4408
E-ISSN: 2347-4734

paper. First, the fuzzy and possibilistic c-means (FCM and PCM) clustering algorithms are analyzed and some drawbacks and limitations are pointed out. Second, based on the reformulation theorem, by means of modifying PCM model, an effective and efficient clustering algorithm is proposed here, which is referred to as a modified PCM clustering (MPCM).



5.3. Map Reduce:

Map Reduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. The model is a specialization of the split-apply-combine strategy for data analysis. Here Hardtop platform with implemented the certain process as finding overall data with mapping and how much data will be reduced. The term Map Reduce actually refers to two separate and distinct tasks that Hardtop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples.

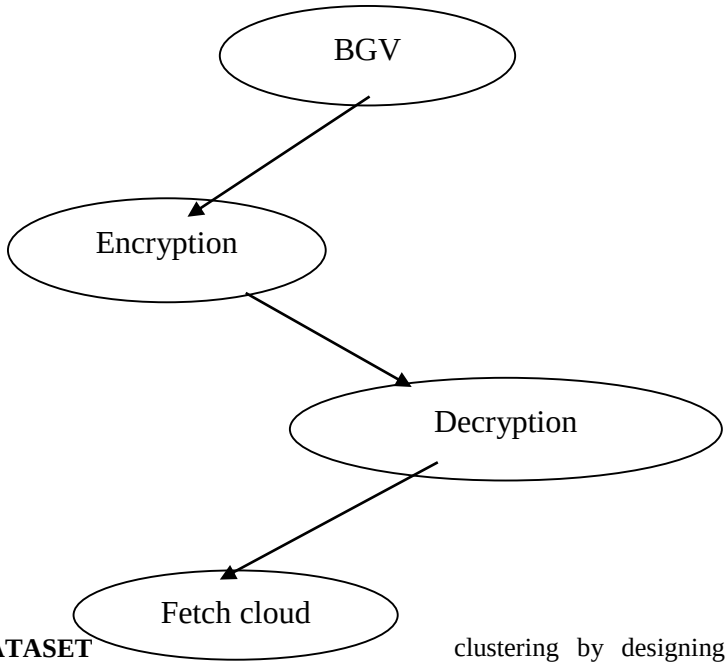


5.4. Fetch Cloud:

Fetch cloud is the extracting data from the cloud server through some security mechanism. Most cloud storage

P-ISSN: 2347-4408
E-ISSN: 2347-4734

providers support Web architectures based on representational state transfer (REST) application programming interfaces (APIs).Some also support traditional block- and file-based data, and cloud storage gateway providers can help customers access data in major storage clouds. To aid the clustering process in this task, we performed pre-processing steps such as feature selection and Principal Component Analysis (PCA) and still, the choice of clustering method is not a trivial one. To find the best performing algorithm.



6. EXPERIMENTAL DATASET

To evaluate the performance of the proposed algorithm carry out some experiments on three representative multimedia data set: NUS-WIDE, CUAVE, and SNAE. The NUS-WIDE data set largest web image set, consists of 269 648 images. Each dataset consists of 1000 images falls into 14 categories. The CUAVE dataset composed of 36 individuals saying the digits 0and 9.Tha CUAVE dataset generate three different data subset i.e.., an image-text subset, an image-audio subset and a text-audio subset. The SNAE collected a total of 180 video clips to form a SNAE dataset from YOUTUBE. The video dataset is classified into four clusters: sports, news, advertisement and entertainment.

7. CONCLUSION

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. One property of the paper is to use the BGV technique to develop a privacy-preserving HOPCM algorithm for preserving privacy on cloud. Experimental results show PPHOPCM can cluster big data by using the cloud computing technology without disclosing privacy. In fact, for the large scale of heterogeneous data that does not require to be protected, the DHOPCM is more suitable since it is more efficient than PPHOPCM. The efficiency of PPHOPCM and DHOPCM can be further improved when using more cloud servers, making them more suitable for big data clustering, since they are of high scalability demonstrated by the experimental results. In this work, the proposed schemes are preliminarily evaluated on two representative heterogeneous datasets. In the future work, the proposed algorithms will be further validated on larger actual datasets.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

REFERENCES:

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ”Data Mining with Big Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.

[2] B. Ermis, E. Acar, and A. T. Cemgil, ”Link Prediction in Heterogeneous Data via Generalized Coupled Tensor Factorization,”Data Mining and Knowledge Discovery, vol. 29, no. 1, pp. 203-236, 2015.

[3] Q. Zhang, L. T. Yang, and Z. Chen, ”Deep Computation Model for Unsupervised Feature Learning on Big Data,”IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016.

[4] 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.

[5] Z. Xie, S. Wang, and F. L. Chung, ”An Enhanced Possibilistic c-Means Clustering Algorithm EPCM,”Soft Computing, vol. 12, no. 6, pp. 593-611, 2008.

[6] Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, ”An Incre-mental CFS Algorithm for Clustering Large Data in Industrial Internet of Things,” IEEE Transactions on Industrial Informatics, 2015. DOI: 10.1109/TII.2017.2684807.

[7] X. Zhang, ”Convex Discriminative Multitask Clustering,” IEEE Transac-tions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 28-40, Jan. 2015.

[8] B. Gao, T. Liu, T. Qin, X. Zheng, Q. Cheng, and W. Ma, ”Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts,” inProceedings of the 13th Annual ACM International Conference on Multimedia, 2005, 112-121.

[9] Y. Chen, L. Wang, and M. Dong, ”Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering,”IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459-1474, Oct. 2010.

[10] L. Meng, 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.

[11] 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.

[12] R. Krishnapuram and J. M. Keller, ”A Possibilistic Approach to Clustering,” IEEE Transactions on Fuzzy Systems, vol. 1, no. 2, pp. 98-110, May 1993.

[13] R. Krishnapuram and J. M. Keller, ”The Possibilistic c-Means Algorithm: Insights and Recommendations,” IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, pp. 385-393, Aug. 1996.

[14] Q. Zhang and Z. Chen, ”A Weighted Kernel Possibilistic c-Means Algo-rithm Based on Cloud Computing For Clustering Big Data,” International Journal of Communication Systems, vol. 27, no. 9, pp. 1378-1391, 2014.

[15] Q. Zhang, 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.

[16] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, ”A Possibilistic Fuzzy c-Means Clustering Algorithm,”IEEE Transactions on Fuzzy Systems, vol. 13, no. 4, pp. 517-530, Aug. 2005.

[17] M. Yang and C. Lai, ”A Robust Automatic Merging Possibilistic Clustering Method,” IEEE Transactions on Fuzzy Systems, vol. 19, no. 1, pp. 26-41, Feb. 2011.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

[18] 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.

[19] A. Schneider, ”Weighted Possibilistic c-Means Clustering Algorithms,” in Proceedings of the 9th IEEE International Conference on Fuzzy Systems, 2000, pp. 176-180.

[20] B. Liu, S. Xia, Y. Zhou, and X. Han, ”A Sample-Weighted Possibilistic Fuzzy Clustering Algorithm,”Acta Electronica Sinica, vol. 30, no. 2, pp. 371-375, 2012.

R. Zhao and W. Grosky, ”Narrowing the Semantic Gap Improved Text-Based Web Document Retrieval Using [21]Visual Features,”IEEE Transactions on Multimedia, vol. 4, no. 2, pp. 189-200, Jun. 2002.

[22]T. Jiang and A.-H. Tan, ”Learning Image-Text Associations,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 2, pp. 161-177, Feb. 2009.

[23]M. Rege, M. Dong, and J. Hua, ”Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering,” in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 317-326.

[24]X. Cai, F. Nie, H. Huang, and F. Kamangar, ”Heterogeneous Image Feature Integration via Multi-Modal Spectral Clustering,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1977-1984.

[25]B. Long, 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.

[26]Q. Gu and J. Zhou, ”Co-Clustering on Manifolds,” in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 359-367.

[27]R. Bekkerman, M. Sahami, and E. Learned-Miller, ”Combinatorial Markov Random Fields,” in Proceedings of the 17th European Conference on Machine Learning, 2006, pp. 30-41.

[28]L. Kuang, F. Hao, L. T. Yang, M. Lin, C. Luo, and G. Min, ”A Tensor-based Approach for Big Data Representation and Dimensionality Reduction,”IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 280-291, Sept. 2014.

[29]Y. Liu, Y. Liu, and K. Chan, ”Tensor Distance based Multilinear Locality-preserved Maximum Information Embedding,” IEEE Transactions on Neural Network, vol. 21, no. 11, pp. 1848-1854, Nov. 2010.

[30]J. Dean and S. Ghemawat, ”Map Reduce: Simplified Data Processing on Large Clusters, “Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

J. Yuan and S. Yu,”Privacy Preserving Back-propagation Neural Net-work Learning Made Pact.