DEVELOPMENT OF FINE-GRAINED BIG DATA PRIVACY PRESERVING VSSF ALGORITHM FOR COST MINIMIZATION

[1]R.Santhana Krishnan, [2]Dr.K.Sasikala, [3]Dr.S.Senthilkumar
[1]Post Graduate Student, [2]Associate Professor, [3]Assistant Professor
[1]Department of IT,
[1]Vinayaka Mission's Kirupananda Variyar Engineering College,
[1]Vinayaka Mission's Research Foundation (Deemed to be university), Salem, India
[2]Department of CSE,
[2]Vinayaka Mission's Kirupananda Variyar Engineering College,
[2]Vinayaka Mission's Research Foundation (Deemed to be university), Salem, India
[3]Department of CSE,
[3]Vinayaka Mission's Kirupananda Variyar Engineering College,
[3]Vinayaka Mission's Research Foundation (Deemed to be university), Salem, India

Abstract: Unfathomable quantum of comprehensive private data is habitually gathered as the mutual exchange of the corresponding information has come as a shot in arm for a multitude of data mining applications. The related data extensively encompass the shopping trends, criminal records, medical history, credit records and so forth. It is true that the corresponding information has proved its mettle as a vital asset to the business entities and governmental organization for the purpose of taking prompt and perfect decisions by means of assessing the pertinent records. However, it has to be borne in mind that harsh privacy. The Big data processing, in fact, involves the explosive expansion of demands on evaluation, storage, and transmission in data centers, thus leading to incredible working expenses to be borne by the data center providers. Thus, the issue of cutting down the expenses has emerged as the most vital factor for the imminent big data era. Here, we explain the PLATFORA algorithm to design the big data processing for high data delivery. The utility-based privacy preservation has two objectives: ensuring the private data and protecting the information utility however much as could be expected. Moreover, protection conservation is a hard prerequisite, that is, it must be fulfilled, and utility is the measure to be optimized. To achieve this, we introduce VSSFA and Map Reduce Framework in Cloud environment. In this proposed work develop a privacy preserving clustering process with cost minimization for big data processing.

***Keywords*: Radial Basis Function, Variation Step Size Firefly Algorithm, Feed Forward Neural Network, Probabilistic Clustering Algorithm.**

I. INTRODUCTION

Data over the internet has been rapidly increasing day by day. Automatically mine useful information from the massive data has been a common concern for the organizations having large dataset. Here, the privacy preserving is one of the larger concerns. Still, it is highly beneficial, as it is home to huge-volume, multifaceted and emergent data sets with multiple, autonomous sources. Regulations and the parallel secrecy constraint are likely to have a dissuading influence on the data owners from parting with the data for effective data appraisal. There is no second opinion that Privacy is the most vital quality which has to be accorded top priority by any smart information system. Thanks to this phenomenon, a host of efforts have been dedicated for integrating the privacy preserving methods with data mining approaches so as to avert the revelation of susceptible data in the course of the knowledge discovery.

With a view to mutually exchange the data with the concurrent conservation of confidential data, the owner has to offer an appropriate solution which is competent to realize the twin object of both privacy preservation and the precise clustering outcome. The leading challenge in this regard is the effective recognition of the clusters in multi-dimensional data sets. Further, it is also plagued by the open challenges in regard to secrecy and safety aspects. With an eye on effectively addressing the corresponding thorny issues, in this document, an earnest endeavor is made to kick-start a novel clustering Probabilistic Possibility Fuzzy C Means Clustering (PFCM) approach viz. the A priori enhancement algorithm for effectively mining the frequent closed item sets. The multidimensional data sets are competent to offer the superlative quality of clustering as needed by the user in view of the fact that datasets can be effortlessly accumulated from several domains. In the golden age of big data, the assessment and extraction of data from the titanic data sets has emerged as a very daunting challenge. The Big data appraisal has illustrated its enormous skills in revealing precious insights of data which has gone a long way in scaling up the efficiency of decision making, considerably cutting back the risks and designing novel products and services. However, it is unfortunate; the big data has transformed itself into a cost-prohibitive venture

thanks to the large amount of expenses and outlays needed for the evaluation and communication, thereby eating into the scarce resources of the companies concerned.

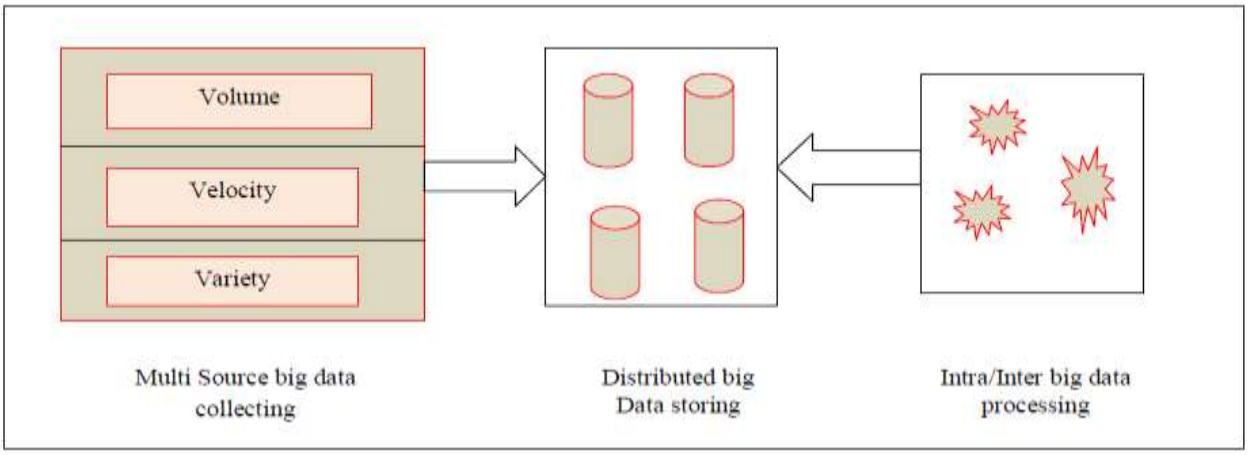

Figure1: General architecture of big data

Of late, the privacy preserving in data mining, cloud computing and big data has surfaced as a dynamic investigation domain with a host of varied applications. Figure 1 explain general architecture for big data. There is no second opinion on the fact that privacy is one of the vital qualities which a data system has to satisfy. In view of this, a number of investigations have been carried out which are dedicated to the incorporation of privacy preserving approaches so as to avert the exposure of sensitive data in the course of the data discovery. In the cloud, the customers are equipped for deflecting colossal capital speculation of the IT base, and center their consideration all alone center business. Consequently various organizations or associations have been moving or incorporating their business with the cloud. Nevertheless, a number of potential clients continue to be hesitant to exploit the full advantage of cloud on account of the safety aspects 18. In addition, the extraction of data from a voluminous quantity of data is a challenging issue in the data mining systems. In this manner, privacy is a standout amongst the most imperative properties that a data framework must fulfill. For this reason, the privacy preserving algorithm is utilized for this research. Chapter 2 introduces the literature survey of Privacy preserving process with Big Data Processing and chapter 3 explains Privacy Preserving proposed algorithm and it details. Chapter 4 explains the result and explanation and chapter 5 explains conclusion for proposed system.

II. LITERATURE REVIEW

In this section we explained the big data classification. Among the big data classification Pekka Paakkonen and Daniel Pakkala proficiently designed the [1] reference architecture and classification of technologies, products and services for big data systems. A supplementary contribution was the classification of corresponding execution techniques and technologies and products/services, which was dependent on the assessment of the published use instances and survey of the associated work. Similarly, Isaac Triguero *et al.* [2] intelligently tabled the Map Reduce Solution for Prototype Reduction in Big Data Classification. The technique aimed at characterizing the original training data sets as an abridged number of instances. Their vital objective was to accelerate the classification procedure and cut down the storage requisites and susceptibility to noise of the nearest neighbour rule.

Moreover, Cem Tekin and Mihaela van der Schaar [3] convincingly discussed the Distributed Online Big Data Classification Using Context Information. In their distributed online data classification structure, data was collected by the distributed data sources and processed by a varied set of distributed learners who pursued learning online, during the execution period , regarding the classification of diverse data streams either by employing the locally accessible classification tasks or by mutual assistance by carrying out classification the data of each other. Additionally, Junchang Xin *et al.* [4] excellently flagged off the Elastic extreme learning machine for big data classification. They launching of the Elastic ELM based on Map Reduce framework which initially evaluated the transitional matrix multiplications of the rationalized training data subset, and thereafter modernized the matrix multiplications by optimizing the previous matrix multiplications with the transitional ones. In [5], Victoria Lopez *et al.* victoriously launched the Cost-sensitive linguistic fuzzy rule based classification systems under the Map Reduce framework for imbalanced big data. Their innovative technique elegantly employed the Map Reduce structure to allocate the computational functions of the fuzzy model which contained the cost-conscious learning methods in its design to successfully tackle the inequity which was abundant in the data. A spreading activation technique of spatial big data retrieval in accordance with the spatial ontology model was effectively envisioned by Shengtao Sun *et al.* [6], in which they explained the speedy growth of spatial data, conventional cause-effect appraisal and conditional recovery deficiency in the age of big data. In their document, they also furnished certain active examples and introduced a prototype.

The Data Mining with the Big Data was fantastically flagged off by XindongWu *et al.* [7]. In their document, they deeply discussed the complex to growing data sets with manifold concerns, large-volume and self-governing sources. They efficiently offered a HACE concept which represented the features of the revolution of Big Data. In an identical way, Xingjian Li [8] excellently elucidated the Mining Frequent Itemsets from Library Big Data. With the intention of tackling the challenge, they brilliantly brought in a superior FP-Growth technique, to which they offered the fond name _RFP-Growth' for the purpose of keeping at bay the creation of intra-property frequent itemsets.

Kaitai Liang *et al.* [9] have explained the Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage. It consolidates the benefits of intermediary re-encryption with unknown method in which a ciphertext was safely and restrictively shared various times without releasing both the learning of hidden message and the character data of ciphertext senders/beneficiaries. Moreover, Yuri Demchenko *et al.* [10] deftly discussed the Instructional Model for Building effective Big Data Curricula for Online and Campus Education. In the document, they were able to furnish the background data on the modern proposals and actions associated with the data swap and synchronization on the design of the educational materials and programs on the Big Data, Data Science, and Research Data Management.

R.Sreedhar and D.Umamaheshwari [11] scientifically designed the Big-Data Processing with Privacy Preserving Map-Reduce Cloud. At the outset, the technique constructively performed the Map Reduce on cloud to Top-Down specialization (TDS) for data anonymization and intentionally devised a group of innovative Map Reduce assignment to effectively execute the specializations in an incredibly scalable manner. Moreover, Xuyun Zhang *et al.* [12] excellently brought to limelight the Proximity-Aware Local-Recoding Anonymization with Map Reduce for Scalable Big Data Privacy Preservation in Cloud. They intelligently introduced a proximity privacy technique which enabled the semantic proximity of sensitive values and multiple sensitive attributes, and transformed the challenge of local recoding as a proximity-aware clustering issue.

Similarly, Chhaya *et al.* [13] charismatically designed the Privacy Preservation Enriched Map Reduce for Hadoop Based Big Data Applications, and the corresponding techniques included the privacy characterization model, anonymizer for datasets, dataset update and privacy preserved data management. The innovative method empowered the data users with the skills to regain the datasets in its unrevealed versions which facilitates the user task dispensing with the need for publishing vital detail particulars regarding the original data. Moreover, They first formulate the general architecture of big data analytics, identify the corresponding privacy requirements, and introduce an efficient and privacy-preserving cosine similarity computing protocol as an example in response to data mining's efficiency and privacy requirements in the big data era. Moreover, Mehdi Sookhak *et al.* have explained the securing big data storage in cloud.

III. PROPOSED SYSTEM

The basic idea of our research is to privacy preserving big data through VSSFA and Map reduce framework in cloud environment. Here, the big data set $D_1$is divided into the number of sub dataset D_1^s, D_2^s, D_3^s. After that in each dataset we apply the two modules such as conditional entropy module and classifier based utility measure using radial basis function-neural network (RBF-NN) module finally Map reduce framework module. Overall diagram of the proposed framework is shows in figure2.A convolution process is applied to the dataset with the help of the variation step size firefly algorithm VSSFA and the output of the privacy dataset is given to the RBF-NN. The implementation is done using JAVA and the performance of the proposed algorithm is compared with existing work algorithm for the benchmark datasets.

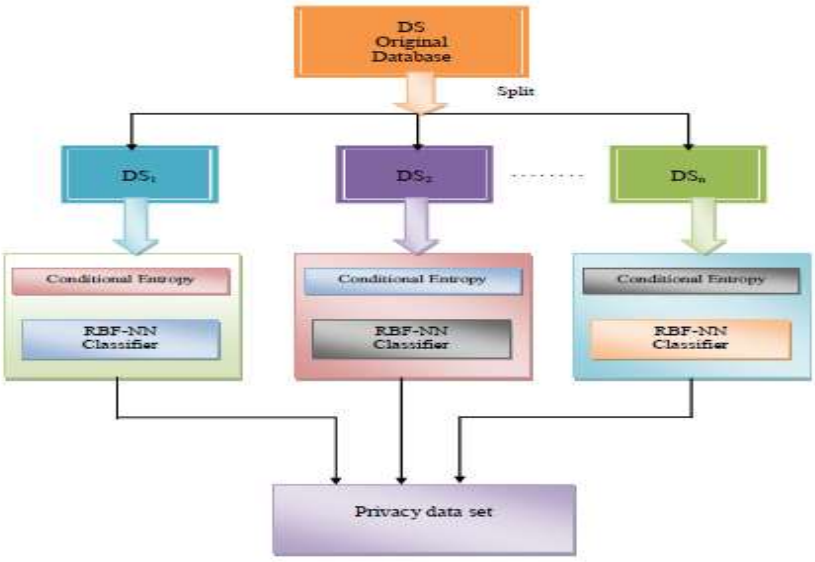

Figure 2 Overall diagram of the proposed framework

In standard FA, the method of setting step α is static. It cannot really reflect the searching process. In general, it is useful for fireflies to explore new search space with a large step, but it is not helpful to the convergence of global optimum. If step has a small value, the result is contrary. Therefore, the step α has a great affect on the exploration and convergence of the algorithm. It would be beneficial to balance the ability of global exploration and local exploitation, and it should also be concerned with its current situation. For this reason, we design a dynamic adjusting scheme of step α which can be controlled. In this paper, we use a non- linear Equation and design a dynamic adjusting scheme of step. The step α can be calculated as following

Step 1: Generate initial solution of fireflies *X* arbitrarily.
$X=\{x_1, x_2, x_3 \ldots x_n\}$
Step 2: Estimate the brightness of the firefly by means of the objective function
$B=\{b_1, b_2, b_3 \ldots b_n\}$
Step 3: set light absorption coefficient γ
Step 4: while (t< max imum iteration)
For i=1 to n
For j=1 to i
If($i_j > i_i$)
Move firefly *i* to firefly *j* by using equation (3).
End if
Attractiveness varies with distance via γ via ($-\gamma r^2$) Evaluate new fireflies and update brightness
End for
End for
t=t+1
end while
Step-5: Rank fireflies according to their fitness and find the best one.
Step-6: If Stopping criteria is reached, then go to step-7.
Else go to step-4.
Setp-7: Stop.

Figure 3: Pseudo code for Variation step size firefly algorithm

The algorithm discontinues its execution only if maximum number of iterations is achieved and the fireflies which are holding the best fitness value is selected. Once the convolution process is done, two phases of privacy-persevering framework over big data in cloud systems is performed. In the first phase, an efficient classifier based utility measure is developed using radial basis function-neural network (RBF-NN), which should capture the intrinsic factors that affect the quality of data for our application. In every iteration, we measure the database different ratio to improve the accuracy of privacy data. In the database difference ratio is maximum we stop the iteration. Based on the output of the RBF-neural network we adjust the conditional entropy output.

IV.RESULT AND DISCUSSION

This section presents the results obtained from the experimentation and its detailed discussion about the results. The proposed approach of Privacy Preserving over Big Data through VSSFA and Map Reduce Framework in Cloud environment is experimented with the data set *Census-Income* (KDD) [35] and UCI machinery Adult dataset. We have implemented our algorithm using Java (jdk 1.6) with cloud Sim tools and a series of experiments were performed on a PC with Windows 7 Operating system at 2 GHz dual core PC machine with 4 GB main memory running a 64-bit version of Windows 2007. In our experimentation we utilized the data set Census-Income (KDD) . This dataset having 299285 records and 40 attributes.

a) Performance analysis based on accuracy:

Here, the performance of the approach is explained based on the accuracy of the privacy preserving data. Here, we use different types of dataset and different methods to prove the accuracy of our approach. The following figures 4.a and 4.b show the performance of the proposed approach using accuracy measures.

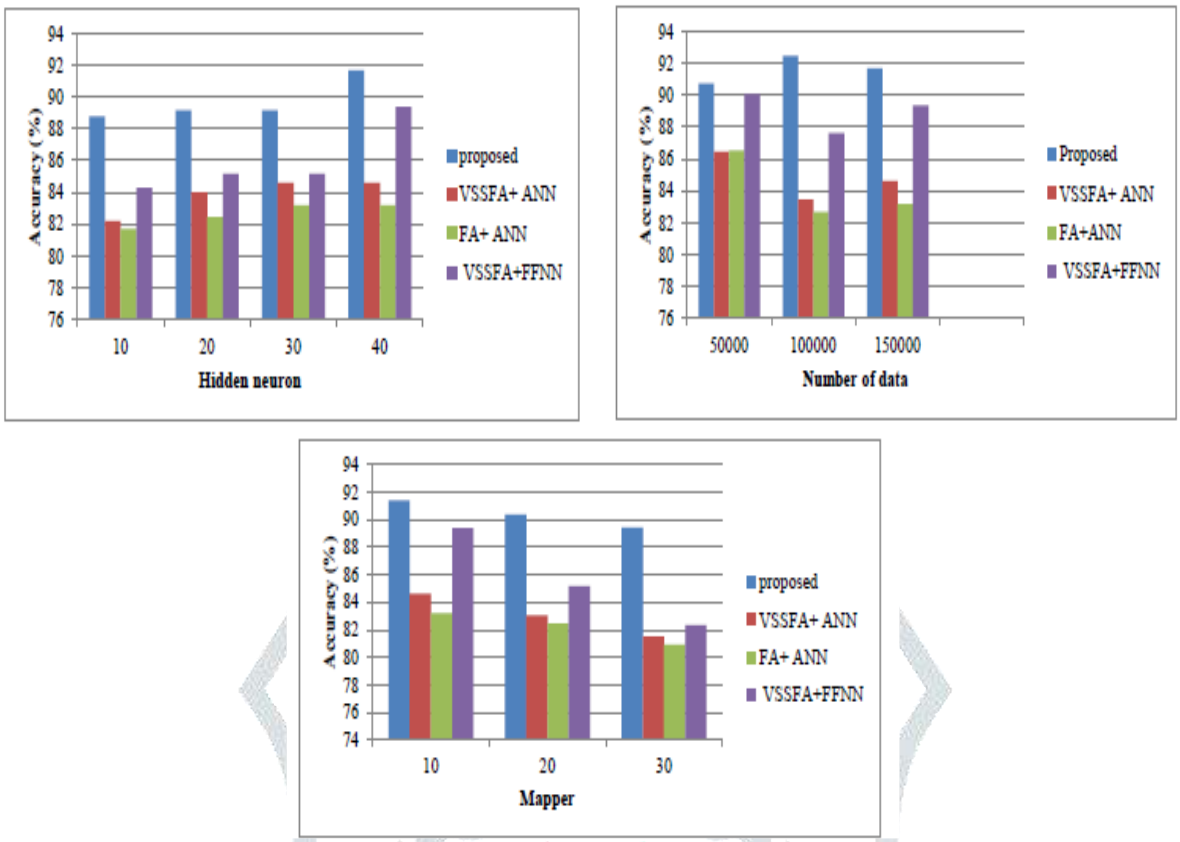

Figure 4: a) performance analysis of accuracy plot of Census-Income (KDD) dataset by hidden neuron size,
b) accuracy plot of Census-Income (KDD) dataset by varying number of data
c) Accuracy plot of Census-Income (KDD) dataset by varying Mapper

The basic idea of our research is to design and develop a technique for Privacy Preserving over Big Data through VSSFA and Map Reduce Framework in Cloud environment. Here, at first we take conditional entropy to the database with the help of the VSSFA optimization algorithm. To improve the privacy of the dataset we further using the RBF based neural network classifier. Finally we map the dataset which produce the privacy data. Here, we prove our work efficiency we compare our work with different approaches such as VSSFA+ANN, VSSFA+FFNN and FA+ANN. Moreover in figure 4, C shows the performance analysis of accuracy plot of Census-Income (KDD) dataset by varying mapper. In the mapper size is 10 our proposed approach achieves the maximum accuracy of 90.34%, which is 81.49% for using VSSFA+ANN , 80.94% for using FA+ANN and 82.34% for using VSSFA+FFNN. When we using the mapper size are 30 we obtain the maximum accuracy of 91.37% which is 84.61% for using VSSFA+ANN, 83.18% for using FA+ANN and 89.37% for using VSSFA+FFNN. From the figure we clearly understand our proposed approach achieves the maximum accuracy compare to existing approaches.

b) Performance analysis based on execution time:

In this section we explained the performance of the proposed approach based privacy preserving based on execution time.

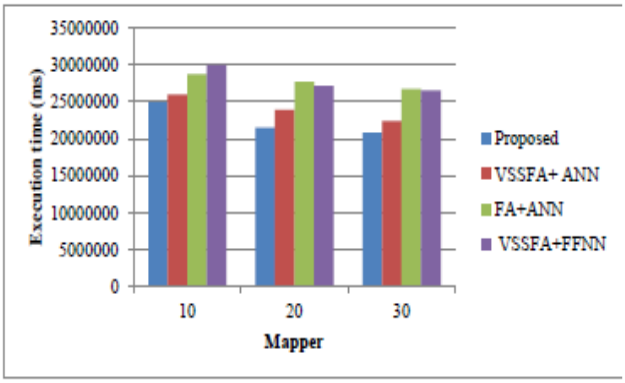

Figure 5: Performance analysis of execution time of Census-Income (KDD) dataset based on Mapper

Moreover, in figure 5 shows the Performance analysis of execution time of Census-Income (KDD) dataset based on Mapper. Here, in our proposed privacy preserving we obtain the minimum execution time of 20846516ms. Consider the all the method the execution time is minimum for our proposed approach.

c) Performance analysis based on the memory usage:

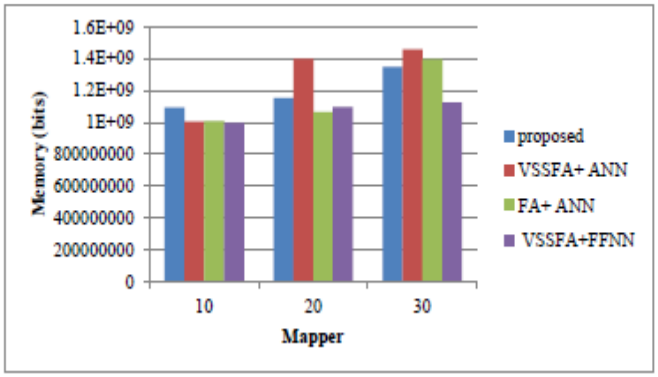

Figure 6: performance analysis of memory usage of Census-Income (KDD) dataset based on mapper

Moreover in figure 6 shows the performance analysis of memory usage of Census-Income (KDD) dataset based on mapper. Here, VSSFA+ANN approach using the maximum memory of 1457654169 bits which is high compare to our proposed work. From the figure, our proposed work using the maximum memory of 1349549747 bits.

Table 1: Performance analysis of our approach using database different ratio

Approaches	Data base different ratio
Proposed	91.05
VSSFA+ ANN	85.64
FA+ANN	84.36
VSSFA+FFNN	88.96

V.CONCLUSION

The current proposal offers a divergent technique of effectively employing the privacy preserving clustering procedure with added emphasize on the incredible cost reduction for gigantic data processing. In this regard, proposed system vibrant and proficient methods are kick-started devoted for the purpose of total privacy preservation. The optimal conditional entropy is taken using the variation step size firefly VSSFA algorithm. The convolution process is used to improve the privacy of the data. Finally the privacy data is given to the RBF-NN, Which is improving the accuracy of the privacy data. Finally, the implementation is done using JAVA and the performance of the algorithm will be analyzed with benchmark dataset. As per the experimentation the proposed algorithm achieves the maximum accuracy compare to the existing approaches.

REFERENCES

[1] Cem Tekin and Mihaela van der Schaar, ―Distributed Online Big Data Classification Using Context Information‖,
[2] Junchang Xin, Zhiqiong Wang, Luxuan Qu and Guoren Wang, ―Elastic extreme learning machine for big data classification‖, Neuro computing, vol. 149, pp.464–471, 2015
[3] Victoria Lopez, Sara Del Rio, Jose Manuel Benitez and Francisco Herrera, ―Cost-sensitive linguistic fuzzy rule based classification systems under the Map Reduce framework for imbalanced big data‖, Fuzzy Sets and Systems, vol. 258, pp.5-38, 2015
[4] Shengtao Sun, Jibing Gong, Jijun He and Siwei Peng, ―A spreading activation algorithm of spatial big data retrieval based on the spatial ontology model‖, Springer Science and Business Media New York, pp. 1-19, 2015.
[5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, ―Data Mining with Big Data‖, IEEE Transactions on Knowledge and Data Engineering vol. 26, no. 1, 2014.
[6] Xingjian Li, "An Algorithm for Mining Frequent Itemsets from Library Big Data", journal of software, vol. 9, no. 9, 2014.

[7] Kaitai Liang, Susilo and Liu, "Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage",Information Forensics and Security, IEEE Transactions on, vol.10, no. 8, 2015

[8] Yuri Demchenko, Emanuel Gruengard and Sander Klous, ―Instructional Model for Building effective Big Data Curricula for Online and Campus Education‖, International Conference on Cloud Computing Technology and Science, 2014

[9] R.Sreedhar and D.Umamaheshwari, ―Big-Data Processing with Privacy Preserving Map-Reduce Cloud‖, International Journal of Innovative Research in Science, Engineering and Technology, vol.3, no.1, 2014.

[10] Xuyun Zhang, Wanchun Dou, Jian Pei, Surya Nepal, Chi Yang, Chang Liu and Jinjun Chen, "Proximity-Aware Local-Recoding Anonymization with Map Reduce for Scalable Big Data Privacy Preservation in Cloud‖, IEEE transactions on computers, 2013.

[11] Chhaya S Dule,H.A. Girijamma and K.M Rajasekharaiah, "Privacy Preservation Enriched MapReduce for Hadoop Based BigData Applications", American International Journal of Research in Science, Technology, Engineering & Mathematics,2014.

[12] Rongxing Lu, Hui Zhu; Ximeng Liu; Liu and J.K.Jun Shao,―Toward efficient and privacy-preserving computing in big data era‖, IEEE Communication socity, vol.28, no.4, 2014.

[13] Mehdi Sookhak, Abdullah Gani, Muhammad Khurram Khan and Rajkumar Buyya, ―Dynamic remote data auditing for securing big data storage in cloud computing‖, Information Sciences, 2015.