DOI: http://dx.doi.org/10.26483/ijarcs.v9i6.6348

ISSN No. 0976-5697

Volume 9, No. 6, November-December 2018

International Journal of Advanced Research in Computer Science

SURVEY PAPER

Available Online at www.ijarcs.info

EFFICIENT DATA MINING TECHNIQUES FOR BIG DATA ANALYSIS: A SURVEY

M.Amsaveni And S.Duraisamy
PG and Research Department of Computer Science
Chikkanna Govt Arts College Tirupur,Tamil Nadu, India.

Abstract: Technology revolution has been facilitating millions of people by generating tremendous data, resulting in big data. It has been a confirmed phenomenon that enormous amount of data have been generated continuously at unprecedented and ever increasing scales. Even though, big data bears greater value, it brings tremendous challenges to extract hidden knowledge and more valuable insights from big data. The valuable information in big data can be obtained by applying data mining techniques in big data. The goal of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between data. Big data mining techniques involves various process like feature selection, clustering and classification. In this article, a detailed comparative survey on different processes of big data mining techniques such as dimensionality reduction, clustering and classification for big data analysis is presented. At first, different dimensionality reduction, clustering and classification methods proposed for big data analysis in previous researches are studied in detail. After that, a comparative and state-of-the-art analysis is carried out to identify the limitations in those methods.

Keywords: Big data, Data mining, dimensionality reduction, clustering, classification.

1. INTRODUCTION

In digital world, data are generated from different homogenous and heterogeneous sources and the fast transition from digital technologies has led to growth of big data [1]. It provides evolutionary breakthroughs in various fields with collection of large datasets. In general, big data refers to the collection of large datasets and complex datasets which are difficult to process using traditional database management tools or data processing applications. Analysis of these massive amounts of data requires a lot of efforts at multiple levels to extract knowledge for decision making.

Data mining refers to the process of searching, analyzing and extracting valuable required data from database to exploit problem-solving and decision making. It involves various processes like pre-processing, feature selection, clustering and classification. These processes can be used over big data to extract knowledge for decision making. Due to the advent of big data feature selection [2] has a key role in helping reduce high dimensionality problems. Clustering [3] is an essential data mining process for analyzing big data. As big data is referring to terabytes and petabytes of data and clustering algorithms are come with high computational costs. In order to find meaningful and accurate data from large unstructured data is dreary task for any users. This is the reason why classification techniques [4] came into picture for big data. With the help of classification methods unstructured data can be turned into organized form so that a user can access the required data easily. In this article, a comprehensive and state-of-the art survey on the feature selection on big data, big data clustering and big data classification is presented. Initially, the most important methods for feature selection, clustering and classification on big data are reviewed in detail. Then, the advantages and the shortcomings of each method are discussed in such a way their limitations encouraged to further improvement on those methods.

2. SURVEY ON BIG DATA ANALYSIS TECHNIQUES

2.1 Survey on Dimensionality Reduction Techniques in Big Data

A Hybrid Genetic Algorithm with Wrapper-Embedded feature approach (HGAWE) [5] was proposed for feature selection in big data analysis. This approach combined the global search and local search by integrating the genetic algorithm with embedded regularization approach. In addition to this, a novel chromosome representation was proposed for local and global optimization procedures in HGAWE. According to the chromosome representation, the regularization method was selected the relevant features in the big data. Simultaneously a learning model was constructed. In order to optimize the control parameters in non-convex regularization, the genetic operations were used.

A hybrid approach called Ant Colony Optimization- Artificial Neural Network (ACO-ANN) [6] was proposed for feature selection in big data environment. ACO algorithm was used to evaluate the selection process and the ANN was used as the classification in ACO-ANN approach. ACO reduced the dimensionality of original data through selection of optimal features by updating position and velocity of each ant in the population. The selected features were used in ANN which classified the best subset from all subset of features and categorized the text.

A novel lightweight feature selection called Accelerated Particle Swarm Optimization (APSO) search feature selection [7] was proposed for data stream mining big data. The process of APSO is same as the PSO which selects the optimal features based on the intensity of each particle in the population. However, the starting positions of PSO must be set appropriately for better feature selection. This was achieved by APSO. The ideal starting positions for APSO were found by using Clustering Coefficients of Variations (CCV). It found a subset of features useful for optimally balancing the classification model induction between generalization and overfitting. The selected features by APSO were used in both traditional and incremental classifier to classify the data stream mining.

A new processing approach [8] was proposed for cancer gene prediction. This approach was structured based on feature extraction and selection. Based on correlation and rank analysis the feature extraction was processed which reduced the number of variables in gene data. Then the redundant variables in the gene data were removed by feature selection approach which using the process of Linear Discriminative Analysis (LDA). It selected the features based on the prediction of the dependent variable value of data.

A novel framework [9] was presented which combined distributed feature selection approach and econometric models for efficient economic big data analysis. A subtractive clustering based feature selection algorithm was developed to identify the important attributes in the economic data. Subtractive clustering is a density-based clustering algorithm which investigated the correlation between data samples. Then it was integrated with attribute coordination to identify the representative attributes. These feature selection processes combined with the econometric model construction to capture the hidden patterns for economic development.

A feature selection algorithm called MapReduce for Evolutionary Feature Selection (MR-EFS) [10] was presented based on evolution computation that used MapReduce paradigm for big data classification. A MapReduce algorithm was developed in such a way that, it divided the original data and performed a group of EFS processes in the map phase and then combined the solutions in the reduce phase. It allowed a flexible application of the feature selection procedure using a threshold to determine the selected subset of features. The selected features were applied in three different classifiers are Support Vector Machine (SVM), Logistic Regression (LR) and Naïve Bayes (NB) for big data classification.

A holistic approach [11] approach was presented to distributed dimensionality reduction of big data. In this approach, a chunk tensor method was presented which fused the structured, semi-structured and unstructured data as a unified model in which all characteristics of the heterogeneous data were appropriately arranged along the tensor orders. A Lanczos based High Order Singular Value Decomposition algorithm was proposed to reduce the dimensionality of the unified model. A Transparent Computing paradigm and linear predictive model were employed to construct the distributed computing platform and to partition the data blocks respectively. It executed the dimensionality reduction task effectively.

The dimensionality reduction methods described in the above section is analyzed and compared based on methods used, their merits, demerits and the parameters used in experimental results. The comparison is given in Table 1.

Table 1. Comparison based on Dimensionality Reduction Methods for Big data analysis

Ref No.	Methods Used	Merits	Demerits	Performance Metrics
[5]	Hybrid Genetic Algorithm with Wrapper-Embedded feature approach	Identify more relevant genes accurately and efficiently	Genetic algorithm is sensitive to the initial population used	AML dataset: Testing Accuracy= 97.84% Training Accuracy= 94.32% DLBCL dataset: Testing Accuracy= 97.28% Training Accuracy= 93.73% Lymphoma dataset: Testing Accuracy= 98.51% Training Accuracy= 94.03% Prostate dataset: Testing Accuracy= 98.32% Training Accuracy= 94.17% Lung cancer dataset: Testing Accuracy= 98.83% Training Accuracy= 93.61%
[6]	Ant Colony Optimization-Artificial Neural	Efficient and optimal for text categorization feature	Low accuracy	Reuters' dataset: Accuracy = 81.35$\pm$2.0 Precision (Acquisition) = 90.52

	Network	selection		Recall (Acquisition) = 92.87
[7]	Accelerated Particle Swarm Optimization	Enhanced analytical accuracy with reasonable processing time	For Naïve Bayes classifier, PSO has better accuracy than APSO	UCI dataset: PSO: Average Accuracy (Traditional Classifier) = 0.29 Average Accuracy (Incremental Classifier) = 0.63 APSO: Average Accuracy (Traditional Classifier) = 0.35 Average Accuracy (Incremental Classifier) = 0.79
[8]	Linear Discriminative Analysis	Good classification	Needs improvement for multiple class prediction	Leukemia dataset: Accuracy = 98% Prostate tumor dataset: Accuracy = 95.5% SRBCTs dataset: Mean accuracy = 90%
[9]	Subtractive Clustering, attribute coordination, economic model construction	Distills the hidden relations	High computational complexity	Nil
[10]	MapReduce for Evolutionary Feature Selection	Better scalability	Threshold value highly influence the classification accuracy	Eplison dataset: Execution Time = 6531 secs Area Under the Curve (AUC) (@1000 features) = 0.737
[11]	Lanczos based High Order Singular Value Decomposition algorithm	Efficient for dimensionality rediction of big data	Can provide a low rank approximation for the initial tensor which is not the best approximation of the initial data	Approximation Ratio (@12 experiments) = 7% Reduction Ratio (@12 experiments) = 86%

2.2 Survey on Clustering Techniques in Big Data

A new ensemble method called fuzzy c-means and cluster ensemble with random projection [12] was presented for big data clustering. The ensemble method was based on partition on similarity graph. For each random projection process, a new data set was generated. The membership matrices were obtained after performing FCM clustering on the new datasets. The elements of membership matrices were treated as similarity measures between points and cluster centers. The spectral embedding of data points were obtained by applying Singular Value Decomposition (SVD) on the concatenation of membership matrices.

An efficient Fuzzy C-means approach [13] was proposed based on tensor canonical polyadic decomposition for big data clustering. In this approach, the conventional fuzzy c-means clustering was converted to the tensor format through a bijection function so that the canonical polyadic decomposition can compress the attributes. In addition to this, the tensor canonical polyadic decomposition was utilized to minimize the attributes of every object before loading the dataset into the memory. The fuzzy c-means method was extended to a high-order fuzzy c-means method to make the clustering operations executed on the compressed objects in the tensor space.

A new distributed clustering approach [14] was proposed for big data clustering. It efficiently dealt with two phases are generation of local results and generation of global models by aggregation. In the first phase of this approach, analyzed the datasets located in each site using K-means and DBSCAN clustering techniques. Then in the second phase of the clustering approach, aggregation phase was designed in such a way that the final clusters were compact and accurate while the overall process is efficient in memory and time allocation. One of the key outputs of this distributed clustering technique was dynamic and there is no need to be fixed in advance.

A novel approach was proposed [15] for improved clustering results in gene expression big datasets. This approach was based on Interval Type-2 fuzzy uncertainty modeling. Initially, a gene expression data was collected as matrix. Then the gene expression data was converted into interval type-2 fuzzified data by using a membership function generation process. Then a crisp equivalent of the fuzzified dataset was obtained by applying an efficient Improved Nie-Tan defuzzification method. Then the defuzzified data were clustered using Fuzzy C Means clustering (FCM).

A modified K-means algorithm [16] was proposed for big data clustering. The selection of initial centroids in k-means algorithm greatly influences the time consumption and complexity of clustering process. Moreover, it changes in data clusters in the subsequence iterations. After a certain number of iterations a small part of the data points changes their clusters. The modified K-means algorithm found the initial centroids and created an interval between those data elements which will not change their cluster in the subsequence iterations. Hence, it minimized the workload significantly in case of big datasets.

A secure weighted possibilistic c-Means algorithm (SWPCM) [17] was proposed for big data clustering. This algorithm was proposed based on Brakerski, Gentry and Vaikuntanathan (BGV) encryption scheme which was utilized to encrypt the raw data for privacy preservation on the cloud. A Taylor theorem was employed to approximate the functions for calculating the weight value and updating the membership matrix. In order to perform correctly and securely on the encrypted data, calculated the cluster centers as the polynomial functions which only included multiplication and addition operations which is named as weighted possibilistic c-Means algorithm.

The big data clustering methods described in the above section is analyzed and compared based on methods used, their merits, demerits and the parameters used in experimental results. The comparison is given in Table 2.

Table 2. Comparison based on Big data Clustering methods for Big data Analysis

Ref No.	Methods Used	Merits	Demerits	Performance Metrics
[12]	Fuzzy c-means and cluster ensemble with random projection	Have more robust partition solutions	There is no proper explanation about how to choose proper number of random projections for cluster ensemble method	ACT2 data set: Fuzzy Rand Index (@ 100 dimension) = 0.86 Xie-Bein Index (@100 dimension) = 0.5 Rand Index (@100 dimension) = 0.9245
[13]	Efficient fuzzy C-means approach	Enhance the cluster efficiency	FCM is affected by the initialization	eGSAD dataset: Adjusted Rand Index (@8 Rank) = 0.71 E_*(@8 Rank) =26.94
[14]	Distributed clustering approach	No need to set the number of global clusters in advance	The quality of clustering depends heavily on the local clustering used during the first phase	Convex type dataset: Execution time (@14000 size) = 290 ms Execution time (@17080 size) = 337 ms Execution time (@30350 size) = 501 ms

[15]	Novel approach	More efficient clustering results for uncertain gene expression dataset	It does not works well with large datasets	LD50-FOU dataset: SIlhouette Index (@3 clusters) = 0.4 Cluster Validity Index(@3 clusters) = 0.47 LD70-FOU dataset: SIlhouette Index (@3 clusters) = 0.45 Cluster Validity Index(@3 clusters) = 1 RD50-FOU dataset: SIlhouette Index (@3 clusters) = 0.41 Cluster Validity Index(@3 clusters) = 1 RD70-FOU dataset: SIlhouette Index (@3 clusters) = 0.43 Cluster Validity Index(@3 clusters) = 1
[16]	Modified K-means	Solve the selection of initial cluster problem effectively	High execution time	Random dataset: Execution time (@5k points) = 35.36 secs Execution time (@10k points) = 92.88 secs Execution time (@50k points) = 405.80 secs Execution time (@5k points) = 667.67 secs
[17]	secure weighted possibilistic c-Means algorithm	Good scalability	Low drop of clustering accuracy by SWPCM	eGSAD dataset: C_* = 12.16 ARI (U, U^*) = 0.91 sWSN dataset: C_* = 0.51 ARI (U, U^*) = 0.87

2.3 Survey on Classification Techniques in Big Data

A MapReduce based distributed framework called MapReduce- Extreme Learning Machine (MR-ELM) [18] was proposed for big data classification. More specifically, MR-ELM was designed for real-world cloud environment in which the huge volume of sample blocks were located in different nodes of hadoop cluster and these were accessed by hadoop file system. With the help of MapReduce framework, training was moved to hadoop nodes which contributed to costs few I/O and high parallelism. ELM sub models were trained parallel with the distributed data blocks on the cluster and then combined as a complete single hidden layer feed forward neural network.

A scalable and distributed dendritic cell algorithm [19] was proposed for big data classification. Dendritic cell algorithm is a bio-inspired classifier which was improved by distributed dendritic cell algorithm based on the MapReduce framework. This algorithm dealt with high dimensional data sets it appeared mandatory to store all the data in a distributed environment and ensured the computations in a parallel way. Based on this consideration, the entire processes of dendritic cell algorithm were partitioned into elementary tasks and then conquer the intermediate results to finally acquire the better output which was the classes of the antigens.

An Elastic Extreme Learning Machine (E^2LM) [20] was proposed for big data classification. ELM has weak learning ability for the updated large-scale training dataset. This was handled by proposed E^2LM which was developed based on MapReduce framework. Initially, it calculated the intermediate matrix multiplications of the updated training data subset and then updated the matrix multiplications by modifying the old matrix multiplications with the intermediate ones. Then the updtred matrix multiplication was used to obtain the corresponding new output weight vector along with centralized computing. Hence, the efficient learning of rapidly updated massive training dataset was realized effectively.

Cost-sensitive linguistic fuzzy rule based classification system [21] was proposed under MapReduce framework for imbalanced big data classification. The fuzzy rule based classification system had the ability to deal with the uncertainty of data that was introduced in huge volumes of data. This system doesn't adjust the learning in the underrepresented class. This method utilized the MapReduce framework to distribute the computational operations of the fuzzy model while it included cost sensitive learning design in its design to address the imbalance problem in the data.

A new fuzzy rule based classification method called CHI-BD [22] was proposed for big data classification problems. In this method, a new MapReduce solution was provided the same classification performance regardless of the number of mappers used for the execution of big data classification. A new rule for each input sample was generated that allowed one to exploit the full potential of MapReduce. Based on this manner, the learning process was divided into two different stages in order to distribute both the rule generation process and the computation of rule weights.

The big data classification methods described in the above section is analyzed and compared based on methods used, their merits, demerits and the parameters used in experimental results. The comparison is given in Table 3.

Table 3. Comparison based on Big data Classification methods for Big data Analysis

Ref No.	Methods Used	Merits	Demerits	Performance Metrics
[18]	MR-ELM	High speedups	Optimization methods will be used for hidden node combination to achieve the highest generalization performance	Segment benchmark: Accuracy = 0.9412 Delta_ailerons benchmark: Residual Sum of Squares = 0.0002
[19]	Distributed dendritic cell algorithm	Better classification accuracy	Distributed dendritic cell algorithm is sensitive to the input class data order	Area Under Curve = 72.92 F-score = 71.68
[20]	Elastic Extreme Learning Machine	Efficiently learn the rapid updated massive training dataset in bigdata classification	Running time of E^2LM increases when the training data update ratio increases	Running Time (@ 30,00,000 records) = 50 secs Running time (@ 1 slave node) = 300 secs
[21]	Cost-sensitive linguistic fuzzy rule based classification	Handles the imbalanced data effectively	Performance of classification depends on the number of mappers	kddcup dataset: Area Under Curve training (@ 8 mappers) = 0.8753 Area Under Curve testing (@ 8

	system			mappers) = 0.8739 Poker dataset: Area Under Curve training (@ 8 mappers) = 0.6427 Area Under Curve testing (@ 8 mappers) = 0.5478
[22]	CHI-BD	Accuracy does not depends on the classification accuracy	An increase in the data size does not have a linear effect on the execution time	Census dataset: Geometric Mean = 0.5231 Area Under Curve = 0.6220 Higgs dataset: Geometric Mean = 0.5847 Area Under Curve = 0.5848 kdd dataset: Geometric Mean = 0.9937 Area Under Curve = 0.9937 Poker dataset: Geometric Mean = 0.6569 Area Under Curve = 0.6579 Skin dataset: Geometric Mean = 0.9597 Area Under Curve = 0.9605 Susy dataset: Geometric Mean = 0.5524 Area Under Curve = 0.6242

3. CONCLUSION

In this article, a detailed comparative survey on feature selection, clustering and classification on big data for big data analysis is presented. Through this comparative analysis, it is obviously noticed that the previous methods have the objective to extract valuable information from big data through feature selection, clustering and classification process. In this article, feature selection on big data is improved to reduce the complexity for big data classification. Also, big data clustering is improved to enhance the efficiency of big data analysis. Moreover, big data classification is improved to find meaningful and accurate data. This survey also helps in deriving the motivation for our future researches as well.

4. REFERENCES

1. Acharjya, D. P., & Ahmed, K. (2016). A survey on big data analytics: challenges, open research issues and tools. Int. J. Adv. Comput. Sci. Appl, 7(2), 1-11.
2. Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, 86, 33-45.
3. Zerhari, B., Lahcen, A. A., & Mouline, S. (2015, May). Big data clustering: Algorithms and challenges. In Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15).
4. Koturwar, P., Girase, S., & Mukhopadhyay, D. (2015). A survey of classification techniques in the area of big data. arXiv preprint arXiv:1503.07477.
5. Liu, X. Y., Liang, Y., Wang, S., Yang, Z. Y., & Ye, H. S. (2018). A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection. IEEE Access, 6, 22863-22874.
6. Manoj, R. J., Praveena, M. A., & Vijayakumar, K. An ACO–ANN based feature selection algorithm for big data. Cluster Computing, 1-8.
7. Fong, S., Wong, R., & Vasilakos, A. (2016). Accelerated PSO swarm search feature selection for data stream mining big data. IEEE transactions on services computing, (1), 1-1.
8. Badaoui, F., Amar, A., Hassou, L. A., Zoglat, A., & Okou, C. G. (2017). Dimensionality reduction and class prediction algorithm with application to microarray Big Data. Journal of Big Data, 4(1), 32.
9. Zhao, L., Chen, Z., Hu, Y., Min, G., & Jiang, Z. (2018). Distributed feature selection for efficient economic big data analysis. IEEE Transactions on Big Data, (2), 164-176.
10. Peralta, D., del Río, S., Ramírez-Gallego, S., Triguero, I., Benitez, J. M., & Herrera, F. (2015). Evolutionary feature selection for big data classification: A mapreduce approach. Mathematical Problems in Engineering, 2015, 1-11.
11. Kuang, L., Yang, L. T., Chen, J., Hao, F., & Luo, C. (2018). A Holistic Approach for Distributed Dimensionality Reduction of Big Data. IEEE Transactions on Cloud Computing, (2), 506-518.
12. Ye, M., Liu, W., Wei, J., & Hu, X. (2016). Fuzzy-means and cluster ensemble with random projection for big data clustering. Mathematical Problems in Engineering, 2016.

13. Bu, F. (2018). An efficient fuzzy c-means approach based on canonical polyadic decomposition for clustering big data in IoT. Future Generation Computer Systems.
14. Bendechache, M., Kechadi, M. T., & Le-Khac, N. A. (2016, October). Efficient large scale clustering based on data partitioning. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), (pp. 612-621).
15. Shukla, A. K., & Muhuri, P. K. (2019). Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets. Engineering Applications of Artificial Intelligence, 77, 268-282.
16. Fahad, S. A., & Alam, M. M. (2016). A modified K-means algorithm for big data clustering. International Journal of Science, Engineering and Computer Technology, 6(4), 129.
17. Zhang, Q., Yang, L. T., Castiglione, A., Chen, Z., & Li, P. (2018). Secure weighted possibilistic c-means algorithm on cloud for clustering big data. Information Sciences.
18. Chen, J., Chen, H., Wan, X., & Zheng, G. (2016). MR-ELM: a MapReduce-based framework for large-scale ELM training in big data era. Neural Computing and Applications, 27(1), 101-110.
19. Dagdia, Z. C. (2018). A scalable and distributd dendritic cell algorithm for big data classification. Swarm and Evolutionary Computation.
20. Xin, J., Wang, Z., Qu, L., & Wang, G. (2015). Elastic extreme learning machine for big data classification. Neurocomputing, 149, 464-471.
21. López, V., del Río, S., Benítez, J. M., & Herrera, F. (2015). Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets and Systems, 258, 5-38.
22. Elkano, M., Galar, M., Sanz, J., & Bustince, H. (2018). CHI-BD: A fuzzy rule-based classification system for Big Data classification problems. Fuzzy Sets and Systems, 348, 75-101.

