CANONICAL VARIATE FEATURE SELECTION
BASED ADAPTIVE ENHANCED WINNOW MAP
REDUCE CLASSIFICATION FOR PREDICTIVE
ANALYTICS

R. Sivakkolundu¹, V. Kavitha²

¹Hindusthan College of Arts and Science, Coimbatore, Tamilnadu, India
²Hindusthan College of Arts and Science, Coimbatore, Tamilnadu, India
sivakkolundu.stud@gmail.com

Abstract. Classification is a key issue to be resolved in data mining. Few research works have been designed for performing predictive analysis through classifying the information on data warehouse. But, classification accuracy (CA) of conventional works was lower when considering a big size of data as input. In order to addresses this drawback, a Canonical Variate Feature Selection based Adaptive Enhanced Winnow Map Reduce Classification (CVFS-AEWMRC) Method is proposed. The CVFS-AEWMRC Method is designed for organizing and classifying the collected and stored data for decision making. Initially, Canonical Variate Feature Selection (CVFS) is carried out in CVFS-AEWMRC Method to select the relevant features for performing the classification. Canonical Variates analysis is a machine learning technique used to find linear combinations of features which have maximum correlation with each other. The features with maximum correlation are selected for performing the classification. Then, Adaptive Enhanced Winnow Map Reduce Classification (AEWMRC) Process is carried out in CVFS-AEWMRC Method to classify the stored data for taking decision. Adaptive Enhanced Winnow technique learns the linear classifier from labeled data samples. Winnow employs the multiplicative scheme for performing the classification process. Winnow learns the hyperplane to classify the data points for decision making. By this way, the data classification is carried out in accurate manner for decision making during the predictive analytics process. Experimental analysis of CVFS-AEWMRC Method is performed on metrics namely feature selection rate (FSR), CA, classification time (CT) and False positive rate (FPR) with number of features and data points.

Keywords: Canonical Variate Feature Selection, Correlation, Decision Making, Features, MapReduce, Prediction Rule, Predictive Analytics

1 INTRODUCTION

In data analysis, historical data storing is significant. A data warehouse is employed in recent decades to perform data analysis and is the basis for analytic reports, data min-
ing and business intelligence which is significant for decision making in organizations, companies, etc. The volume of data from different sources increasing in an exponential pattern complicates data storage and retrieval in data warehouse, results in bottlenecks. In big data, data warehouse management is a prominent issue due to the high rate of data growth and the need for data analysis. Hence a novel method called CVFS-AEWMRC is introduced in this research work.

In [1], Label-Aware Distributed Ensemble Learning (LADEL) was designed to handle big data. However, FSR using LADEL was higher. An Atrak method using Map Reduce-based data warehouse was presented in [2] for big data to enhance query execution speed through employing node independence. But, CA of predictive analytics was not sufficient.

An extension of Technology–Organization–Environment (TOE) framework was presented in [3] for discover barriers and opportunities of Artificial Intelligence (AI). However, the feature selection process was not carried out in efficient manner. To address the data warehouse issues, A new Chabok method was introduced in [4] which utilize two-phased Map-Reduce. Chabok not required the data replication for the join omission.

High-Level MapReduce Query Languages was employed in [5] directly on MR which translated queries to executable native MR jobs. But, MapReduce-based HLQL failed to reduce the computational cost. A semi-automatic design methodology was introduced in [6] for decision making. However, FPR of predictive analytics was higher.

Boolean query generator utilizing common-table expressions was employed in [7] for giving better clinical data warehouse performance. But computational complexity was more. To design big data warehouse, an efficient and real-time power monitoring platform with open source techniques was introduced in [8]. However, time complexity using this algorithm was not reduced.

Hadoop as big data analytic was implemented in [9] with data warehouse to support decision making process. But predictive analytics performance was poor. A novel method was used in [10] to addresses the data warehouses issues and fulfilling requirements of decision-makers. However, FPR using this method was not reduced.

The characteristics of collected and stored data are utilized in this article to construct a novel method called CVFS-AEWMRC for predictive analysis.
• Initially, CVFS method is employed in CVFS-AEWMRC method to choose the relevant features. After that, an AEWMRC is designed in CVFS-AEWMRC for decision making.

• In data warehouse management system, Potential information is predicted using CVFS-AEWMRC method which assists in decision makers for efficient predictive analytics.

2 LITERATURE SURVEY

By considering the distributed storage and parallel processing, a large-scale system was developed in [11] for social media data warehousing with the assist of Hadoop and Storm. A simple low-overhead mechanism was introduced in [12] to evaluate every data portion for efficient big data processing in warehouse-scale computers.

In [13], Game Theory based Materialized View selection (GTMV) was designed to select optimal set of views for materialization and thus increases the data warehouse performance. Specification of the data warehouse was employed in [14] for the decision-making dimension of the Bid Process Information System.

Parallel Processing System was employed in [15] to increase the capability of processing data on warehouse. A novel methodology was developed in [16] for designing data warehousing applications from different sources.

A mixed approach was introduced in [17] for creating XML data warehouses. An integrated optimization method was used in [18] to provide better computational method for distributed data warehouses.

An extract transforms and load strategy were introduced in [19] for unstructured data into data warehouse with map reduce paradigm for big data analytics. With the help of uniform distributed dataset, a novel method was designed in [20] to resolving data warehouse issues in big data.

3 CANONICAL VARIATE FEATURE SELECTION BASED ADAPTIVE ENHANCED WINNOW MAP REDUCE CLASSIFICATION METHOD

The CVFS-AEWMRC Method is designed with aim of performing an efficient data classification for accurate decision making with a minimal time. On the contrary to conventional works, CVFS-AEWMRC Method is proposed by combining the CVFS algorithm and AEWMRC algorithm. The designed CVFS algorithm in CVFS-AEWMRC Method helps to enhance the feature selection performance for effective
predictive analytics. Besides to that, AEWMRC algorithm in CVFS-AEWMRC Method supports for classifying the collected and stored data on warehouse by using selected features and thereby enhances CA of predictive analytics.

Figure 1 presents the flow processes of CVFS-AEWMRC Method. From figure 1, CVFS-AEWMRC Method considers Instacart Market Basket Analysis dataset as input. Followed by, CVFS-AEWMRC Method chooses key features using CVFS algorithm. In CVFS algorithm, Canonical correlation analysis is utilized to discover and compute the associations among two sets of variables for selecting relevant features from big dataset. With the selected relevant features, the collected and stored data are classified for decision making. The detailed processes of CVFS-AEWMRC Method are described in below sections.

3.1 Canonical Variate Feature Selection

CVFS algorithm is designed in CVFS-AEWMRC Method to attain better FSR for predictive analytics. CVFS algorithm is proposed to identify and determine the asso-
ciations among two sets of variables. Besides, CVFS algorithm finds orthogonal linear combinations of the variables within each feature that best explain the variability both within and between features. This helps for CVFS algorithm to extract significant features for accurate predictive analytics.

Figure 2 depicts the flow processes of CVFS algorithm to carry out feature selection process. As demonstrated in figure, the association among two features is computed by Canonical correlation analysis in CVFS algorithm. Let assume two features ‘\( a_1 \)' and ‘\( a_2 \)', and there are correlations between the information, and Canonical correlation analysis find out linear combinations of \( a_1 \) and \( a_2 \) which have a higher correlation with each other. Accordingly, canonical correlation is mathematically estimated as,

\[
\alpha \rightarrow \sum a_1 a_2 = Cov(a_1, a_2)
\]  \hspace{1cm} (1)

From the above mathematical formula (1), the covariance matrix is evaluated with help of information in features ‘\( a_1 \)' and ‘\( a_2 \)’. Followed by, the covariance among two features ‘\( a_1 \)' and ‘\( a_2 \)' is mathematically obtained as,
From the above mathematical representation (2), ‘\(a_1\)’ signifies the values of the data in features ‘\(a_1\)’ and ‘\(a_2\)’ designates the values of data in features ‘\(a_2\)’. Here, ‘\(\bar{a}_1\)’ point outs the mean of the features ‘\(a_1\)’ and ‘\(\bar{a}_2\)’ indicates mean of the features ‘\(a_2\)’ and ‘\(n\)’ indicates a total number of features in big dataset. By using equation (1) and (2), CVFS algorithm calculates the correlation between the two features. Thus, CVFS algorithm extracts the features with higher correlation value as significant for accurately performing data classification process.

\[
Cov(a_1, a_2) = \frac{\sum(a_1 - \bar{a}_1)(a_2 - \bar{a}_2)}{n}
\]  

(2)

The process of CVFS is described in algorithm 1. With the help of above algorithmic steps, CVFS-AEWMRC method accurately selects relevant features in a given dataset for accomplishing predictive analytics process with higher accuracy. Thus, CVFS-AEWMRC method improves the FSR of predictive analytics when compared to existing works.
3.2 Adaptive Enhanced Winnow Map Reduce Classification

In CVFS-AEWMRC method, AEWMRC algorithm is designed for performing decision making with enhanced data classification performance with minimal time. AEWMRC algorithm is a machine learning technique. To perform decision making with less time consumption, AEWMRC algorithm is utilized which categorizes the collected and stored data. The AEWMRC is related to the perceptron algorithm. The perceptron algorithm in conventional works utilizes additive weight-update scheme. In contrast to this, AEWMRC applies a multiplicative scheme which helps to perform better data classification process while many dimensions are unrelated (thus it’s called winnow). The designed AEWMRC algorithm in CVFS-AEWMRC method is works well while taking a large size of dataset. On the contrary to existing works, AEWMRC algorithm finds a decision hyperplane with help of MapReduce function that used to classify all collected and stored data. AEWMRC algorithm process is shown in below Figure 3.

![Figure 3 Flow processes of AEWMRC algorithm](image)

**Figure 3 Flow processes of AEWMRC algorithm**

Figure 3 shows the overall processes of AEWMRC algorithm. As presented in the above figure, AEWMRC algorithm initially takes a number of collected and stored data ‘$d_1, d_2, \ldots, d_n$’ in warehouse as input where ‘$n$’ signifies a total number of
The instance space is $d = \{0, 1\} \times \{0, 1\}^n$ i.e. each input data in AEWMRC algorithm is represented as a set of Boolean-valued features. Subsequently AEWMRC give non-negative weights $w_i$ for each data $i \in \{1, \ldots, n\}$ which are primarily set to ‘1’, one weight for each data $i \in \{1, \ldots, n\}$. Then, AEWMRC algorithm measures the similarity between each data and prediction rule using below,

$$s = \frac{n \cdot \sum d_i \cdot p_r}{\sum d_i^2 + \sum p_r^2 - \sum d_i \cdot p_r}$$

(3)

From equation (3), $s$ denotes the tanimoto similarity coefficient, $n$ denotes the number of data, $\sum d_i \cdot p_r$ indicates a sum of product of paired scores, $d_i$ denote collected data, $\sum d_i^2$ represents a sum of squared score of data $d_i$, $\sum p_r^2$ specifies a sum of squared score of prediction rule $p_r$. Prediction rule includes set of conditions to predict the user’s next order. In AEWMRC algorithm, tanimoto similarity value offers output value from 0 to +1. During the mapping process, the AEWMRC algorithm maps all the input data in a big dataset into a corresponding class by using similarity value using below mathematical representation,

$$y^* = \begin{cases} +1 & \text{if } (w^T s_i > b) \\ -1 & \text{otherwise} \end{cases}$$

(4)

From the above formula (4), ‘b’ denotes a threshold assigned for similarity value and ‘$y^*$’ signifies the classified output of AEWMRC algorithm (predicted output). Together with the weight, the threshold determines a separating hyperplane in the instance space to accurately classify collected and stored data in warehouse. Here, ‘+1’ represents that the product is user’s next order whereas ‘-1’ the product is not a user’s next order. During reducing phase, AEWMRC algorithm removes the irrelevant data in each class.

For each data classification result, AEWMRC algorithm employs below update rule:

- If an input data is correctly classified, do nothing.
- If an input data is predicted incorrectly and the correct result was ‘+1’, then the AEWMRC algorithm updates each weight multiplied by ‘2’ i.e. $w_i \leftarrow 2w_i$ only for those data features ‘$d_i$’ that are ‘1’.
- If an input data is predicted incorrectly and the correct result was ‘-1’, then the AEWMRC algorithm updates each weight divided by ‘2’, $w_i \leftarrow w_i/2$ only for those data features ‘$d_i$’ that is ‘-1’.
With the assist of above processes, the AEWMRC algorithm significantly performs the users in making decision with the collected and stored data. The algorithmic processes of AEWMRC are presented in below,

```
// Adaptive Enhanced Winnow Map Reduce Classification Algorithm
Input: Number of collected data \(d_i = d_1, d_2, d_3, ... d_n\); Actual Output \(y\); Predicted Output \(\hat{y}\); weight \(w\);
Output: Achieve higher classification accuracy
Step 1: Begin
Step 2: For each collected data \(d_i\)
Step 3: Initialize weight \(w = 1\)
// Map phase
Step 4: Determine similarity between data and prediction rule using (3)
Step 5: Maps input data into a corresponding class using (4)
// Reduce phase
Step 6: Removes the irrelevant data in each class
Step 7: End for
Step 8: For each obtained classification result
Step 9: If \((y = +1)\) and \((\hat{y} = -1)\) then
Step 10: Update each weight \(\omega_i \leftarrow 2\omega_i\) only for data feature \(d_i\) with ‘1’
Step 11: Else If \((y = -1)\) and \((\hat{y} = +1)\) then
Step 12: Update each weight \(\omega_i \leftarrow \omega_i/2\) only for data feature \(d_i\) with ‘-1’
Step 13: End If
Step 14: Return accurate data classification result
Step 15: End For
Step 16: End
```

**Algorithm 2 Adaptive Enhanced Winnow Map Reduce Classification**

Algorithm 2 shows the step by step processes of AEWMRC. By using the above algorithmic steps, CVFS-AEWMRC method accurately classify the data on data warehouse by using selected features with higher accuracy and lower time. Accordingly, CVFS-AEWMRC method enhances the predictive analytics performance to exactly find user's next order than the conventional works.

## 4 EXPERIMENTAL SETTINGS

To evaluate the performance of CVFS-AEWMRC Method is implemented using MapReduce libraries in Hadoop with assist of Instacart market basket analysis dataset https://www.kaggle.com/c/instacart-market-basket-analysis/data that includes relational set of files describing customers' orders over time. With this data features, CVFS-AEWMRC Method identify the user's next order. This dataset includes 3 mil-
lion grocery samples orders from 200,000 Instacart users. Here, each entity includes of customer, product, order, aisle, etc.

The CVFS-AEWMRC Method considers number of data features ranges from 5000 to 50000. The performance of CVFS-AEWMRC Method is determined in FSR, CA, CT and FPR. The performance result of CVFS-AEWMRC Method is compared against with two conventional methods [1] and [2]. The experimental evaluation of CVFS-AEWMRC Method is conducted and averagely ten results are depicted in below table and graph.

5 RESULTS

The comparative result of CVFS-AEWMRC Method is discussed and comparison is made with existing [1] and [2]. CVFS-AEWMRC Method is determined with the help of tables and graphs.

5.1 Performance Measure of Feature Selection Rate

In CVFS-AEWMRC Method, ‘FSR’ determines the amount of time taken for selecting the significant features. The FSR is mathematically obtained using below,

\[ FSR = n \times \text{time (SSF)} \]  \hspace{1cm} (5)

From equation (5), ‘time (SSF)’ represents the time consumed for choosing one feature, ‘n’ denotes a total number of features taken for experimental evaluation. FSR is calculated in milliseconds (ms).

Table 1 demonstrates the results of FSR obtained during the processes of predictive analytics process with three methods namely proposed CVFS-AEWMRC method and conventional Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2].
### Table 1: Tabulation result of Feature Selection Rate

| No. of features | feature selection rate (ms) | CVTS-AEWMRC | LADEL | Attrak method |
|-----------------|----------------------------|-------------|-------|--------------|
| 5000            | 75                         | 93          | 125   |
| 10000           | 90                         | 100         | 120   |
| 15000           | 105                        | 117         | 128   |
| 20000           | 122                        | 136         | 150   |
| 25000           | 143                        | 173         | 178   |
| 30000           | 150                        | 177         | 195   |
| 35000           | 161                        | 186         | 214   |
| 40000           | 180                        | 200         | 212   |
| 45000           | 183                        | 203         | 225   |
| 50000           | 200                        | 225         | 275   |

#### Figure 4: Graphical Result Analysis of Feature Selection Rate versus Number of Features
FSR results of three methods are portrayed in figure 4 with numbers of features ranges from 5000-50000. As depicted in figure 4, CVFS-AEWMRC method achieves minimal FSR in order to efficiently discover the user’s next order as compared to existing [1] and [2]. This is due to the utilization of CVFS algorithm in CVFS-AEWMRC method.

During the feature selection process, CVFS-AEWMRC method applies canonical correlation analysis determines the correlation between the two features through estimating covariance matrix. By using this analysis, CVFS-AEWMRC method find outs the features with maximum correlation value as optimal with a minimal amount of time consumption. This supports for CVFS-AEWMRC method to minimize the amount of time required for selecting the significant features from an input dataset when compared to existing Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2]. Thus, proposed CVFS-AEWMRC method reduce FSR by 13% and 23% when compared to existing Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2] respectively.

5.2 Performance Measure of Classification Accuracy

In CVFS-AEWMRC Method, ‘CA’ measured as ratio of a number of features which are precisely classified to total number of features. CA is calculated as below,

\[ CA = \frac{\delta_{AC}}{n} \times 100 \]  \hspace{1cm} (6)

From the above mathematical formula (6), ‘\( \delta_{AC} \)’ represents the number of features accurately classified in which ‘\( n \)’ symbolizes a total number of features employed for experimental process. CA of predictive analytics is evaluated in percentage (%).

Table 2 presents the comparative result of CA acquired during the processes of predictive analytics process using three methods namely proposed CVFS-AEWMRC method and traditional Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2].
Table 2 Tabulation result of Classification Accuracy

| No. of features | Classification Accuracy (%) |
|-----------------|-----------------------------|
|                 | CVFS-AEW Merkel             | LADEL | Attack method |
| 5000            | 98.5                        | 77.2  | 74.8          |
| 10000           | 96.75                       | 83.5  | 84.02         |
| 15000           | 96                          | 82.63 | 76.13         |
| 20000           | 95.43                       | 86.88 | 81.01         |
| 25000           | 95.89                       | 88.9  | 87.96         |
| 30000           | 94.7                        | 90.05 | 89.153        |
| 35000           | 94.41                       | 90.46 | 89.923        |
| 40000           | 95.13                       | 88.73 | 85            |
| 45000           | 94.45                       | 89.05 | 90.71         |
| 50000           | 94.1                        | 89.52 | 86.92         |

Figure 5 Graphical Result Analysis of Classification accuracy versus Number of Features
CA results of three methods are shown in figure 5 with numbers of features ranges from 5000-50000. As demonstrated in figure 5, CVFS-AEWMRC method gets higher CA in order to predict the user's next order as compared to existing [1] and [2]. This is owing to application of CVFS algorithm and AEWMRC algorithm in proposed CVFS-AEWMRC method. This enhances the ratio of number of features correctly classified as compared to conventional [1] and [2]. Accordingly, proposed CVFS-AEWMRC method increase CA of predictive analytics by 10% and 14% when compared to state-of-the-artLabel-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2] respectively.

5.3 Performance Measure of Classification Time

In CVFS-AEWMRC Method, \( CT \) calculates amount of time needed to classify the features for predictive analytics. The \( CT \) is mathematically determined using below,

\[
CT = n \times \text{time (CSF)} \tag{7}
\]

From equation (7), \( \text{time (CSF)} \) indicates time consumed for classify the single feature, \( n \) denotes a total number of features. \( CT \) is estimated in milliseconds (ms).

Table 3 depicts the result analysis of \( CT \) during the predictive analytics process using three methods namely proposed CVFS-AEWMRC method and state-of-the-artLabel-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2].

| No of features | Classification Time (ms) |
|----------------|-------------------------|
|                | CVFS-AEWMRC | LADEL | Atrak method |
| 5000           | 95          | 115   | 150          |
| 10000          | 115         | 130   | 160          |
| 15000          | 135         | 145.5 | 180          |
| 20000          | 154         | 196   | 200          |
| 25000          | 170         | 197.5 | 227.5        |
| 30000          | 180         | 207   | 255          |
| 35000          | 196         | 220.5 | 283.5        |
| 40000          | 220         | 240   | 292          |
| 45000          | 229.5       | 247.5 | 315          |
| 50000          | 250         | 275   | 325          |
CT results of three methods is illustrated in figure 6 with numbers of features ranges from 5000-50000. As exposed in the above graphical figure, CVFS-AEWMRC method obtains lower CT when compared to existing Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2]. This is due to application of CVFS algorithm and AEWMRC algorithm in proposed CVFS-AEWMRC method on the contrary to existing works. This assists for CVFS-AEWMRC method to decrease the amount of time needed to classify the features for predictive analytics when compared to traditional Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2]. For that reason, proposed CVFS-AEWMRC method reduces CT of predictive analytics by 12% and 27% as compared to conventional Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2] respectively.

5.4 Performance Measure of False Positive Rate

‘FPR’ estimated as the ratio of number of features imperfectly classified to total number of features as input. FPR is measured as follows,
\[ \text{FPR} = \frac{\delta_{IC}}{n} \times 100 \]  

(8)

From the above mathematical formulation (8), \( \delta_{IC} \) indicates the number of features wrongly classified in which \( n \) signifies a number of features. The FPR of predictive analytics is calculated in percentage (\%).

Table 4 illustrates the result of FPR during the processes of predictive analytics using three methods namely proposed CVFS-AEWMRC method and existing Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2].

| No. of features | False Positive Rate (%) |
|-----------------|-------------------------|
|                 | CVFS-AEWMRC | LADEL | Atrak method |
| 5000            | 1.5         | 22.8  | 25.2         |
| 10000           | 3.25        | 14.5  | 15.98        |
| 15000           | 4           | 17.37 | 23.87        |
| 20000           | 4.58        | 13.12 | 18.99        |
| 25000           | 4.18        | 11.1  | 12.04        |
| 30000           | 5.3         | 9.95  | 10.85        |
| 35000           | 5.59        | 9.34  | 10.08        |
| 40000           | 4.87        | 11.27 | 14.99        |
| 45000           | 5.55        | 10.95 | 9.29         |
| 50000           | 5.9         | 10.48 | 13.08        |
FPR results of three methods is depicted in figure 7 with numbers of features ranges from 5000-50000. As presented in the above graphical illustration, CVFS-AEWMRC method achieves minimal FPR when compared to traditional Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2]. This is due to the utilization of CVFS algorithm and AEWMRC algorithm in CVFS-AEWMRC method. This aids for CVFS-AEWMRC method to diminish the ratio of a number of features poorly classified for predictive analytics when compared to traditional Label-Aware Distributed Ensemble Learning (LADEL) [1] and Atrak method [2]. Thus, proposed CVFS-AEWMRC method minimize FPR by 61% and 66% as compared to existing [1] and [2].

6 CONCLUSION

The CVFS-AEWMRC method is proposed with the goal of enhancing the decision-making performance for effective predictive analytics through classifying the collected and stored data on data warehouse. The goal of CVFS-AEWMRC method is attained with the support of CVFS and AEWMRC algorithms. The proposed CVFS-AEWMRC method increases the number of features correctly classified with the as-
sist of AEWMRC algorithm than the conventional works. Also, CVFS-AEWMRC method minimize the amount of time consumed for choosing the significant features from an input dataset by using the CVFS and AEWMRC algorithmic steps. Besides, CVFS-AEWMRC method lessens the amount of time needed to classify the features for predictive analytics as compared to conventional works. Moreover, CVFS-AEWMRC method lessen the number of features inaccurately classified for predicting the user’s next order. The experimental result shows that the proposed CVFS-AEWMRC method presents better predictive analytics performance with higher CA and minimum FSR as compared to conventional works.

References

1. Shadi Khalifa, Patrick Martin, Rebecca Young, “Label-Aware Distributed Ensemble Learning: A Simplified Distributed Classifier Training Method for Big Data”, Big Data Research, Elsevier, Volume 15, Pages 1-11, March 2019
2. Mohammadhossein Barkhordari, Mahdi Niamanesh, “Atrak: a MapReduce-based data warehouse for big data”, The Journal of Supercomputing, Springer, Volume 73, Issue 10, Pages 4596–4610, October 2017
3. Kamran Mahroof, “A human-centric perspective exploring the readiness towards smart warehousing: The case of a large retail distribution warehouse”, International Journal of Information Management, Elsevier, Volume 45, Pages 176–190, 2019
4. Mohammadhossein Barkhordari and Mahdi Niamanesh, “Chabok: a Map-Reduce based method to solve data warehouse problems”, Journal of Big Data, Springer, Volume 5, Issue 40, Pages 1-25, 2018
5. Marouane Birjali, Abderrahim Beni-Hssane and Mohammed Erritali, “Evaluation of high-level query languages based on MapReduce in Big Data”, Journal of Big Data, Volume 5, Issue 36, Pages 1-21, 2018
6. Lucile Sautot, Sandro Bimonte, Ludovic Journaux, “A semi-automatic design methodology for (Big) Data Warehouse transforming facts into dimensions”, IEEE Transactions on Knowledge and Data Engineering, June 2019
7. Daniel R. Harris, Darren W. Henderson, Ramakanth Kavuluru, Arnold J. Stromberg, Todd R. Johnson, “Using Common Table Expressions to Build a Scalable Boolean Query Generator for Clinical Data Warehouses”, IEEE Journal of Biomedical and Health Informatics, Volume 18, Issue 5, Pages 1607 – 1613, 2014
8. Chih-Hung Chang, Fuu-Cheng Jiang, Chao-Tung Yang, Sheng-Cang Chou, “On construction of a big data warehouse accessing platform for campus power usages”, Journal of Parallel and Distributed Computing, Elsevier, Volume 133, Pages 40-50, November 2019
9. Leo Willyanto Santoso, Yulia, “Data Warehouse with Big Data Technology for Higher Education”, Procedia Computer Science, Elsevier, Volume 124, Pages 93–99, 2017
10. Ladjel Bellatreche, Sharma Chakravarthy, “A special issue in extending data warehouses to big data analytics”, Distributed and Parallel Databases, Springer, Volume 37, Issue 3, Pages 323–327, September 2019
11. Ferdaous Jenhani, Mohamed Salah Gouider, Lamjed Ben Said, “Streaming Social Media Data Analysis for Events Extraction and Warehousing using Hadoop and Storm: Drug Abuse Case Study”, Procedia Computer Science, Elsevier, Volume 159, Pages 1459-1467, 2019
12. Hossein Ahmadvand, Maziar Goudarzi, “Using Data Variety for Efficient Progressive Big Data Processing in Warehouse-Scale Computers”, IEEE Computer Architecture Letters, Volume 16, Issue 2, Pages 166 – 169, July 2017
13. Hossein Azgomi, Mohammad Karim Sohrabi, “A game theory based framework for materialized view selection in data warehouses”, Engineering Applications of Artificial Intelligence, Elsevier, Volume 71, Pages 125-137, May 2018
14. Manel Zekri, Sahbi Zahaf, Sadok Ben Yahia, “Specification of the data warehouse for the decision-making dimension of the Bid Process Information System”, Procedia Computer Science, Elsevier, Volume 159, Pages 756-763, 2019
15. Fajar Ciputra Daeng Bani, Suharjito, Diana, Abba Suganda Girsang, “Implementation of Database Massively Parallel Processing System to Build Scalability on Process Data Warehouse”, Procedia Computer Science, Elsevier, Volume 135, Pages 68-79, 2018
16. Khouri Selma, Boukhariflyes, Bellatreche Ladjel, Sardet Eric, Jean Stéphane, Baron Mihiel, “Ontology-based structured web data warehouses for sustainable interoperability: requirement methoding, design methodology and tool”, Computers in Industry, Elsevier, Volume 63, Issue 8, Pages 799-812, October 2012
17. Zoubir Ouaret, Doulkifi Boukra, Omar Bousaid, Rachid Chalal, “AuMixDw: Towards an automated hybrid approach for building XML data warehouses”, Data & Knowledge Engineering, Elsevier, Volume 120, Pages 60-82, March 2019
18. P. Amuthabala, R. Santhosh, “Robust analysis and optimization of a novel efficient quality assurance method in data warehousing”, Computers & Electrical Engineering, Elsevier, Volume 74, Pages 233-244, March 2019
19. P. Saravanakumar, M. Athigopal, S. Vetivel, “Extract Transform and Load Strategy for nstructured Data into Data Warehouse Using Map Reduce Paradigm and Big Data Analytics”, International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 12, Pages 7456-7462, December 2014
20. Mohammadhossein Barkhordari and Mahdi Niamanesh, “Aras: A Method with Uniform Distributed Dataset to Solve Data Warehouse Problems for Big Data”, International Journal of Distributed Systems and Technologies (IJDST), Volume 8, Issue 2, Pages 1-14, 2017
21. V. Kavitha, “Clustering Process with Time Series Data Stream”, Smart Innovations, Systems and Technologies, 2020.
22. R. Sivakkolundu, V. Kavitha, “Tobit Regressive Based Gaussian Independence Bayes Map Reduce Classifier on Data Warehouse for Predictive Analytics”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278 – 3075, October 2019, Scopus Indexed.
23. V. Kavitha, “Pre-processed Hierarchical Clustering for Time Series Data Streams”, International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-3, September 2019, Pg.251-254, ISSN: 2277-3878