Abstract

Aim: The main motto of the study is to optimize the large volume of data using data placement algorithm and online community adjustment algorithm and comparing their accuracy. Materials and Methods: Data placement algorithm (N=10) and online community adjustment (N=10) was iterated 20 times to optimize the data. Result and Discussion: Data placement algorithm has significantly better accuracy (85%) compared to online community adjustment algorithm(78%). The statistical significance of data placement (p<0.02 independent sample test) is high. Conclusion: With the limits of the study, a data placement algorithm with product manufacturing data offers the best accuracy in data optimization.

Key-words: Data Placement, Community Adjustment, Innovative Data Optimization, Deep Learning.

1. Introduction

The purpose of this study is to optimize the large volume of data using data placement algorithm and online community adjustment algorithm. To optimize the data for future behaviour in production manufacturing units. Used in manufacturing industry 4.0, and used in smart factories. The data optimization is important for reducing the storage and while the storage is reduced, automatically reduces the storage locations. Using a classification model, the data set will optimize for future behaviour.(Liu et al. 2020)
The temporal pattern classification and analysis of large volumes of data is difficult. So, the outliers and null values are removed in data preprocessing. The huge data helps to get the knowledge about the data and gives the accurate values (Liu et al. 2020). There are 300 papers published on temporal pattern classification and analysis of large volume of data using data mining in sciencedirect and 450 papers on google scholar and 10 papers were published in ieee xplore for fraud detection. (Atrey et al. 2019) et al proposed a cloud space allocation algorithm to optimize data that shows less significance (Atrey et al. 2019). Zhang et al uses cloud computing processes to optimize data. It is a time taken process and obtained accuracy is also low (Zhang et al. 2021).

Previously our team has a rich experience in working on various research projects across multiple disciplines (Sathish and Karthick 2020; Varghese, Ramesh, and Veeraiyan 2019; S. R. Samuel, Acharya, and Rao 2020; Venu, Raju, and Subramani 2019; M. S. Samuel et al. 2019; Venu, Subramani, and Raju 2019; Mehta et al. 2019; Sharma et al. 2019; Malli Sureshbabu et al. 2019; Krishnaswamy et al. 2020; Muthukrishnan et al. 2020; Gheena and Ezhilarasan 2019; Vignesh et al. 2019; Ke et al. 2019; Vijayakumar Jain et al. 2019; Jose, Ajitha, and Subbaiyan 2020). Now the growing trend in this area motivated us to pursue this project.

The existing methods achieve less accuracy and take more time to optimize the data. K. Liu et al. the proposed framework to optimize the data manually with the help of machine learning algorithms and achieves less accuracy (Liu et al. 2020). The main aim of this study is to optimize the large data by using data placement and online community adjustment to compare their accuracy.

2. Materials & Methods

The study setting of the proposed work is done in Saveetha School of Engineering. The number of groups identified for this study is two. Group 1 is given as a data placement algorithm and Group 2 is given as an online community adjustment algorithm. Sample size for each group was calculated by using previous study results in clinical.com by keeping g power as 80 %, threshold 0.05 and confidence interval as 95%(Yu and Pan 2016; Liu et al. 2019). According to that, the sample size of data placement algorithm (N=10) and online community adjustment algorithm (N=10) were calculated.

The dataset is about production manufacturing unit products count. In the manufacturing unit each month they have to manufacture some 500 products. Sometimes from the target 500, they can manufacture more or less products. This information is present in the dataset in the form of dates and

Data Placement Algorithm

The data placement algorithm for distributed storage systems depends on the knowledge of information quality for creating placement selections. The data placement formula provides a reliable storage location. The data programming between the info centers and info acquisition improve effectively ((Yu and Pan 2016; Liu et al. 2019)).

Pseudo Code

Input: Dataset M, node set N, request pattern read rate Rpy, data write rate Wx, master node yx.

Output: Data replica placement εxy, request routing δpyj.

Initialization: ∀εxy ← 0, δpyj ← 1 if j = yx.
1. for Data item x ∈ M, x ∉ Dy do
2. "εxy ← 1, Dy ← x, if Rxy ≥ Wx;
3. end for
4. Exchange the data storage location information Dj with all other nodes, j ∈ N;
5. for Request pattern p ∈ P do
6. Calculate request routing {δpyj} based on Dj, j ∈ N;
7. end for
8. for Data item x ∈ M, x ∉ Dy do
9. Calculate θxpy based on (14), ∀p ∈ Px;
10. εxy ← 1, Dy ← x, if vxy ≤ 0;
11. end for
12. Repeat Step 4 – 7 to update the request routing {δpyj} based on the Dj after the expansion, j ∈ N.

Online Community Adjustment Algorithm

Online community adjustment schemes are proposed to solve the replica placement problem in a scalable and adaptive way. The online scheme is adaptive to handle the bursty data requests. An
online community adjustment scheme is proposed to adaptively handle the bursty requests. Data storage location will be adjusted accordingly for the adaptive community expansion and reduction (Yu and Pan 2016; Liu et al. 2019).

**Pseudo Code**

**Input**: Dataset M, node set N, real-time read/write rate Rtpy and Wtx, master node yx, existing replica placement εxy and request routing δpyj.

**Output**: Updated placement εxy and request routing δpyj.

1. Monitor data item x at node y from t=0 to T, x ∈ M;
2. if $\sum p \in P_x | Rtpy - R_{t\cdot p}y | + | Wty - W_{t\cdot y} | > \phi$ then
3. if εxy = 1 && y ≠ yx && Rtxy < Wtx then
4. εxy ← 0, x ∉ Dy;
5. Calculate $\theta_1xpy$, ∀p ∈ Px;
6. Update routing: $\delta_{pyyx} ← 1$ if $\theta_1xpy = 1$, ∀p ∈ Px;
7. end if
8. if εxy = 0 then
9. Calculate $\theta xp$, ∀p ∈ Px;
10. εxy ← 1, Dy ← x, if vxy ≤ 0;
11. end if
12. If replica x is added/removed at t, node y broadcasts the message εxy to other nodes;
13. end if
14. if Receive the message εxj = 1 or εxj = 0, δpyj = 1 from node j, j ∈ N, j ≠ y then
15. Update request routing {δpyj} with the greedy method in Section 4.2, ∀p ∈ Px;
16. end if.

The software tool used to evaluate the data placement and online community adjustment algorithm was in a jupyter notebook with python programming language. The hardware configuration was intel core i5 processor with a RAM size of 8GB. The system type used was a 64-bit, OS, X64 based processor with HDD of 917GB. The software configuration includes windows 10 operating system.

In the proposed model train the dataset and implement the classification algorithm based on the dataset. After collecting the dataset, the null values and errors were removed. By this the data preprocessing was done. After data preprocessing the dataset is split into two parts one for training
and other for testing. In the dataset 30% is split for training and the remaining 70% given to the testing process. By evaluating the algorithm with train and test sets to perform optimization and achieve better accuracy percentage.

The analysis was done using IBM SPSS version 21. It is a statistical software tool used for data analysis. For both proposed and existing algorithms 10 iterations was done with a maximum of 10-20 samples and for each iteration the optimized accuracy was noted for analysing accuracy. In this research date and name of the product are the independent variables because they are inputs and remain constant even after changing other parameters, whereas pattern and accuracy are dependent variables because they depend on the inputs and vary for every change in the input. The analysis of the research work is done using Independent T-Test which is used to compare data placement algorithm and online community adjustment algorithm to optimize the data.

3. Result

Fig. 1 - Sample entities and attributes of the dataset to optimize large volume of data
Fig. 2 - Results of 583 dataset manufacturing units production values based on expected production and classification production during temporal period of 30 days.

Sample test dataset entities and attributes to optimize large volume of data (Fig 1). Output of training data 553 and testing data 30 are manufacturing units production values, based on this
expected production and classification production during a temporal period of 30 days (Fig 2). 
Comparison of manufacturing units production values based on expected production and 
classification production during temporal period of future 30 days based on data placement algorithm 
and online community adjustment algorithm (Fig 3). The data placement algorithm achieved 
precision 82.7%, recall 65.7%, accuracy 85%, and F-score 86.4%. The online community adjustment 
model achieved 92.4% precision, 92.4% recall, 78% accuracy, and 62.3% F1-score. Finally, the 
proposed classifier achieved an accuracy of 85%. Thus, the model is able to work efficiently in 
temporal pattern classification and analysis of large volume of data using associated data placement 
algorithm and online community adjustment (Table 1). The mean, standard deviation and standard 
error mean of data placement algorithm and online community adjustment based innovative data 
optimization is tabulated. (Table 2) which shows that data placement has an accuracy mean of 
85.38%, Std.Deviation 0.26979 for the sample size of N=10 where the online community adjustment 
has an accuracy mean of 78.04, Std.Deviation of 0.51547 for the sample size of N=10, based on the 
above results the statistical significance of data placement is high. The mean, standard deviation and 
significant difference of data placement algorithm based data optimization and online community 
adjustment based data optimization is tabulated (Table 3) which shows there is a significant 
difference between the two groups since p<0.03 (Independent Sample T Test). The mean, standard 
deviation and standard error mean of online community adjustment based data optimization and data 
placement based optimization is tabulated. Bar graph is comparing the mean accuracy of data 
placement algorithm and online community adjustment algorithm for innovative data optimization 
(Fig 4).

Table 1-optimized accuracy to detect frauds (online community adjustment accuracy of 78% and data placement algorithm 
accuracy of 85%)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>F1 score</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Community Adjustment</td>
<td>78%</td>
<td>62.3%</td>
<td>92.4%</td>
<td>92.4%</td>
</tr>
<tr>
<td>Data Placement Algorithm</td>
<td>85%</td>
<td>86.4%</td>
<td>65.7%</td>
<td>82.7%</td>
</tr>
</tbody>
</table>

Table 2 - Group statistics results (mean of data placement 85.389 is more compared with online community adjustment 
78.04 and std.Error Mean for DPA is 0.08532 and OACA is 0.16301.).

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std.Deviation</th>
<th>Std.Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DPA</td>
<td>10</td>
<td>85.3890</td>
<td>0.26979</td>
<td>0.08532</td>
</tr>
<tr>
<td></td>
<td>OACA</td>
<td>10</td>
<td>78.0400</td>
<td>0.51547</td>
<td>0.16301</td>
</tr>
</tbody>
</table>
Table 3 - Independent sample T-test Results is applied for a dataset fixing confidence interval as 95% and level of significance as 0.02(Data placement appears to perform significantly better than online community adjustment with the value of p=0.020).

<table>
<thead>
<tr>
<th>Levene’s test for equality of variances.</th>
<th>T-test for equality of means.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>6.48</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>39.94</td>
</tr>
</tbody>
</table>

Fig. 4 - Comparison of Data Placement algorithm and Online Community Adjustment algorithm in terms of mean accuracy and precision. The mean accuracy and precision of Data Placement is better than Online Community Adjustment. The standard deviation of Data Placement is slightly better than Online Community Adjustment. X Axis: Data Placement vs Online Community Adjustment. Y Axis: Mean accuracy of detection ± 1 SD

4. Discussion

Data placement algorithm based data optimization has better accuracy compared to online community adjustment algorithm based data optimization from large volumes of data.

K. Liu et.al has implemented a data placement algorithm and online adjustment algorithm to optimise the data from the cloud and obtained 65% accuracy (Liu et al. 2020). Artery et al introduced the data mining models and they used to optimize the data for future behaviour and obtained 70% accuracy (Atrey et al. 2018).
The factors that affect the data optimizations are computational cost, null values, data type mismatch and dataset size. The identification ability of the model is completely dependent on the data size and its characteristics; small size datasets with a smaller number of class labels performs better convergence. The research is aimed to develop simple networks to reduce the storage locations (Zhou et al. 2016) these networks produce good results against large data sets. Some simple pre-trained networks have found difficulty in learning one class successfully with high accuracy. K.Liu et.al have proposed online community adjustment algorithms to optimise the data from the cloud. Given read and write permission for optimised data from the cloud. So the accuracy percentage is decreased (Liu et al. 2020; Zhou et al. 2016). Charapko et.al. proposed framework to implement the data migration, the topology-aware policies results upto 70%, latency improvement nearly 95% (Charapko, Ailijiang, and Demirbas 2018). There is no opposite finding related to this proposed algorithm.

Our institution is passionate about high quality evidence based research and has excelled in various fields ((Vijayashree Priyadharsini 2019; Ezhilarasan, Apoorva, and Ashok Vardhan 2019; Ramesh et al. 2018; Mathew et al. 2020; Sridharan et al. 2019; Pc, Marimuthu, and Devadoss 2018; Ramadurai et al. 2019). We hope this study adds to this rich legacy.

Due to limitations such as threshold, precision and recall. The production manufacturing units data used in this dataset is collected from various sources. The evaluation of accuracy cannot provide a better outcome on larger data sets. So, the data needs to be optimized. Moreover in online community adjustment, the mean error appears to be higher than data placement. It would be better if the mean error can be reduced to a considerable extent. However, the work can be enhanced by applying innovative data optimization techniques, to achieve a better accuracy and less mean error. Feature optimization algorithms can be used before classification models to improve the classification accuracy of optimize the data.

5. Conclusion

Based on the obtained results the data placement algorithm provides better accuracy (85%) compared to the online community adjustment algorithm provides (78%) accuracy.

Declarations

Conflict of interests: No conflict of interest in this manuscript.
Authors Contributions

Author H. Sudarsan kumar raju was involved in data collection, data analysis, manuscript writing. Author Dr. M. Nalini was involved in conceptualization, guidance and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Es Techno Minds Pvt. Ltd.
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

References


