AEM-DEDUPE- A NOVEL IMPLEMENTATION OF ACTIVE SUPERVISORY FEEDFORWARD NETWORKS FOR DETECTION OF DATA DEDUPLICATION

[1]Mr. N. Lakshmi Narayana,
Research Scholar, Koneru Lakshmaiah Education Foundation, Guntur, India.
thisisnarayan@gmail.com
[2]Dr. B. Tirapathi Reddy,
Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Guntur, India.

Abstract

Development of efficient methods for detection of data deduplication process is interesting as well as challenging in the computing scenario of intensive applications in data, especially in cloud computing. With an advent of machine learning algorithms, challenges in data deduplication process have been reduced to great extent but achieving the higher accuracy of deduplication process still remains in the darker side of the research. This paper presents the novel approach of implementing active feed forward learning models to detect the data deduplication process in the context of digital gazette records. The proposed framework discusses about the extraction of various similarity features such as semantic similarity vectors, time stamp vectors to add the efficiency for the supervised active feed forward learning models. The comprehensive experimentations have been carried out using the different machine learning algorithms and performance metrics such as deduplication accuracy, precision and recall with time complexity were calculated and analyzed. Simulation results shows that the proposed active learning models has outperformed the other learning models which proves more efficient for the data deduplication process.

***Keywords*: Data Deduplication, Machine Learning, Active Feed Forward Learning Models, Similarity vectors, Deduplication Accuracy.**

1. INTRODUCTION:

Data Deduplication in the cloud is an emerging technology that furnishes the exponential growth in data storage for digital data. It identifies the redundant data and eliminates it. The resultant unique data copy will be stored and can furnish them to the authorized users. The process of deduplication is categorized based on the data of user such the single user and cross user data. In the single user data, deduplication is applied to identify the redundancy within his/her data whereas cross -user deduplication is used to identify the redundancy in data among the different users. Cross-user deduplication edges over the Single users in terms of less complexity and less computational cost. According the [1], nearly 60% of data can be deduplicated using only cross -over deduplication process. However, data deduplication process offers more advantages in terms of cloud storage and cloud capacity but still suffers from the less performance in terms of deduplication accuracy and time complexity.

With the advent of machine learning algorithms, data deduplication process has taken its new dimension in improvising the accuracy performance and time complexity. The machine learning based data deduplication process involves the feature extractions methods integrated with the training models. Recent studies reveals that the data deduplication process has inquired into the usage of supervisory machine learning mechanism by employing the trained labelled data into record pairs that can be marked as duplicates or non-duplicates [2,3,4,]. These are known as binary classifiers which identifies the redundancy in data by calculating the various similarity features. Most commonly used machine learning models include the decision trees (DT)[5]- and Support Vector Machines (SVM)[6,7]. The decision tree [8] models are used for the binary classifiers which has tree like model in which the tree models represent the classification problems. Even though the decision tree

offers the best output but it suffers to take decisions intelligently. On the other hand, SVM performs by establishing a hyper-plane that increases the total distance between itself and distinct vector. SVM furnishes the more advantages than the decision tress in terms accuracy of removing the redundancy in the data records. However, the above-mentioned machine learning algorithms fails miserably in evaluating all the possible pairs of duplicate records in the data. To solve these aforementioned problems in the existing machine learning algorithms, this paper proposes the new technique of Active Single Feed forward technique integrated with an effective diversified features to provide an efficient data deduplication mechanism. This research paper presents the development of Single Feedforward Layers for an effective detection of data deduplication process with less time complexity. This kind of learning models can provide the high accuracy in data deduplication process with the less time complexity.

The organization of the paper comprises as: Section-II propounds about the related works of other authors. The working mechanisms, phases of the proposed architecture is presented in section-III. The results, evaluation metrics and analysis are propounded in section-IV. The Section-V concludes with its future scope.

2. RELATED WORKS:

M. Gracious proposed a brought together information deduplication plan with the accompanying attributes. This plan adequately eliminates the unique mark file by utilizing hidden capacity framework's current hash table (Double hashing), and stretch out its metadata design to viably uphold information deduplication (Self-contained item). Likewise, degradation of performance is limited through rate control and specific deduplication dependent on post-preparing. Likewise actualized the proposed plan upon open-source dispersed capacity framework, Ceph. Principle benefit of this strategy is it can uphold high accessibility and superior in conveyed capacity framework: it is relevant on replication or eradication coding and limit execution degradation brought about by deduplication. However, fundamental disadvantage of this structure is overseeing scalability of fingerprint index can be dangerous with a restricted working memory size [9].

Zhu, R presented multi-strung deduplication (multi-Dedup) architecture utilizing equal deduplication strings to conceal the I/O inactivity. A prefix based simultaneous record are intended to keep up the inner consistency of the file that are de-duplicated with lower synchronization overhead. Then again, a collisionless store cluster was likewise intended to protect area and similitude inside the equal strings. In different real world datasets tests, multi-Dedup accomplishes 3−5 time's presentation enhancements fusing with region based ChunkStash and neighborhood closeness-based SiLo techniques. Furthermore, Multi-Dedup has significantly diminished the overhead in synchronization and accomplishes 1.5 to 2 time's exhibition enhancements contrasting with traditional synchronization techniques based on lock. This structure has a advantage of higher adaptability with lower overhead in RAM, yet an imperfection with lower throughput as the information stream has a strong locality [10].

W. Xia propose FastCDC, an effective and fast CDC technique, that forms and enhances the most recent Gear-based CDC approach, one among the quickest CDC techniques as far as anyone is concerned. The critical thought behind FastCDC is the joined utilization of three key procedures, in particular, streamlining and upgrading the hash judgment to treat our noticed difficulties confronting Gear-based CDC, skirting sub-least lump slice highlight further accelerate CDC, and piece size dissemination is normalized in a little determined area to treat the challenges of the diminished deduplication proportion from cut-point skipping. This methodology accomplishes the most noteworthy deduplication proportion however with just a little increasing speed of the lumping speed. In any case, principle constraint is this methodology is delicate to the CPU overhead of substance characterized piecing, for example, QuickSync [11]

D. Bhagwat have presented another strategy, Extreme Binning, for versatile and equal deduplication, which is particularly appropriate for responsibilities comprising of individual records with low region. Extraordinary Binning abuses document similitude rather than territory to make just one plate access for lump query per record rather than per piece, accordingly reducing the circle bottleneck issue. Benefit of this structure is, it parts the chunk list into two levels bringing about a low RAM impression that permits the framework to look after throughput. For a bigger informational collection than a level list plot. Greatest parallelization can be accomplished because of the one document one reinforcement hub conveyance. Reinforcement hubs can be added to support throughput and the rearrangement of lists and lumps is a perfect activity in light of the fact that there are no conditions between the receptacles or between pieces appended to various containers. In any case, the downside is, to accomplish high throughput, reinforcement frameworks need to compose new lumps consecutively to circle. [12].

Q. Zhang Given a portrays the responsibility and machine heterogeneity in Google's creation figure Clusters. At that point introduced a Harmony, a heterogeneity-mindful system that progressively changes the quantity of machines to find some kind of harmony between energy reserve funds and booking delay, in consideration with the reconfiguration cost. The tests are utilized by Google responsibility follows; Harmony yields huge energy reserve funds while altogether improving assignment booking delay. Be that as it may, as CBS receives a severe booking strategy, that is lesser in practical for real-world. Hence, the cloud supplier should cautiously examine these trade-offs to choose which plan ought to be utilized in a given situation [13].

W. Leesakul proposed a powerful information deduplication conspire for distributed storage, to satisfy a harmony between changing stockpiling effectiveness and adaptation to non-critical failure prerequisites, and furthermore to improve execution in distributed storage frameworks. This structure powerfully changes the quantity of duplicates of records as indicated by the changing degree of QoS. The result of the trial show that, this framework is executing admirably and can deal with versatility issue. This plan can result with a negative effect on framework adaptation to non-critical failure. Since there are numerous documents that allude to a similar information piece [14].

Burramukku proposed a mechanism to guarantee secure information deduplication utilizing the upsides of dynamic amazing hash strategies. This system delivers the strategies to give secure deduplication of information, by taking the notoriety of Data things into the count, and accepting that information things require various degrees of security dependent on prominence. Information identified with the prevalence of information things has been kept up by a part of the way confided in key worker that will help the client in knowing which information things are disliked. Yet, this component won't give classification to the information put away when the cloud specialist organization is straightforward however inquisitive [15].

S. Chavhan proposed program consolidates recuperation deleting coding and reinforcement inline derivations in cloud. Eradication helps in encoding the data sets. Assessment of the proposed program is accomplished by exploring different files regarding a few documents in different datasets. The outcome shows that utilization of space is lesser in this Mechanism, along these lines diminishing information misfortune expenses and recuperation. Yet, in this structure execution debased because of extra activities by adding deduplication. The duplication technique is utilized to kill copies from records, subsequently improving capacity frameworks execution. The raw power of a solitary hub is still to a great extent limited to tens or exactly hundred terabytes, prompting complex use [16].

S. Chen proposed the idea of double chunking information deduplication for bringing down the overhead for information with lesser deduplication proportion and guarantee compelling repetition expulsion for information with higher deduplication proportion. To understand the proposed idea, another information deduplication expanded record framework, DeEXT, is proposed. This technique presents a check module that are compressible for foreseeing the deduplication proportion of information streams and applying appropriate chunking technique for every information stream. Also, utilizing a pool-based space allocator for storing deduplicated information piece with little asset pools with information hotness thought. At last, a space recovery conspire is utilized to recover invalid space while adjusting the wearing of every asset pool. This work has huge decrease in inactivity and energy utilization. In any case, limitation is content defined chunking- CDC really checks the information content prior to piecing, the overhead in chunking of CDC is higher than that of fixed-size chunking [17].

P. Hamandawana proposed CROCUS, a structure to empower computed resource orchestration to upgrade group wide deduplication execution. Specifically, this structure considers all figure assets, for example, local as well as remote by overseeing decentralized process pools. An artful ‘Load-Aware Fingerprint Scheduler’, circulates and offloads figure concentrated deduplication activities in a heap mindful style to process pools. Also this structure is exceptionally nonexclusive and can be embraced in both inline and disconnected deduplication with various capacity level setups. Additionally, executed structure in Ceph scale-out capacity framework. Benefit of this system is it can limit the capacity costs by installing metadata the board in each SSD level hub. Yet, there is an intricacy to guarantee similarity between recently applied deduplication metadata and existing metadata. Likewise, deduplication of often utilized information will bring about pointless I/O and possible execution debasement [18].

3. PROPOSED FRAMEWORK:

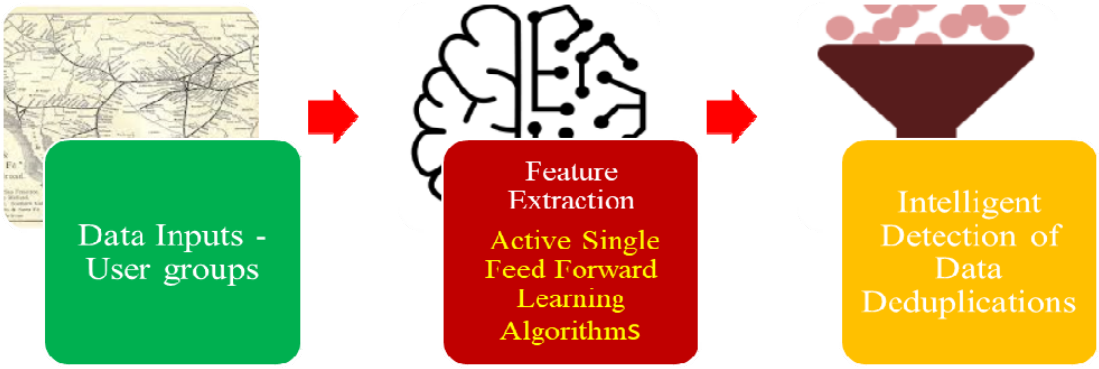

Figure 1 System Framework for the Proposed Data Dedeuplication Architecture

3.1.1 SYSTEM OVERVIEW:

The working mechanism of the proposed architecture is shown in Figure 1. The proposed architecture comprises of the three phases and they are Collection of data, Feature extraction with active single feedforward learning algorithms and intelligent detection of data deduplication. The working principle of each and every stage in the proposed framework is explained below

3.2 DATA COLLECTION PROCESS:

The scenarios for collecting the general dataset for this mechanism are described in this paper and it is one for two gazetteer datasets A and B that are evolved by distinct independent sources. Every record in gazetteer corresponds to the geographic location over the earth's surface. The parameters which is present in the Gazetteer Datasets locations are shown in Figure 2

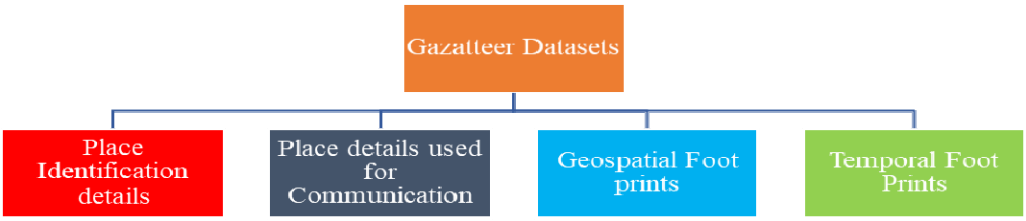

Figure 2 Different Geographic e Location Parameters Present in Gazetteer Datasets

The geographic locations in the gazetteer's datasets are identified by its geographic location of the places, type of places, geospatial footprints which represents the different geometrics located in the earth's surface and temporal footprints which is designed for the temporal interval representation for the place. These gazetteers are created by the geographical markup language (GML) standard. Often, geographic locations in the records are given as multiplicity of distinct place names, different place types and various footprints by several group of people and distinct purposes of time. Hence, within an individual gazetteer, a real-world place and its gazetteer record are standing in a one-to-one relationship. Thus, the construction of the gazetteers created from the multiple sources of users leads to the duplicate replication of the data in various aspects.

Hence these gazetteer record data consist of the duplicate replication as mentioned above in the following ways such that Frequently used same places, misleading information, change of boundaries in the places, sharing of same centroid co-ordinates for the different places and same representation of time zones for different regions of places. The gazetteer records collected for the testing the proposed framework is depicted in table I.

Table I Gazettes Dataset Collection record

Sl.no	Datasets	Characteristics
01	No of records	1,257,00
02	No of Non-duplicated records	1,246,43
03	No of duplicated records	1,957
04	No of places	1,114

3.3 FEATURE EXTRACTION:

After collecting the datasets, feature vectors are calculated based on the similarity between the different metadata present in the gazetteer records. The features used for the detection of data deduplication process are given as follows as 1) Place Similarity Feature (PSF)2) Place Type Similarity Feature (PTSF)3) Geospatial Similarity Feature (GSF) 4) Temporal Similarity Feature (TSF) and 5) Semantic Relationship Feature (SRF). Table II represents the five feature vectors with its characteristics and metrics for extraction.

Table II Different Feature Vectors Used for Data Deduplication Detection Process

Types of the Feature Vectors	Characteristics	Metrics For feature Extraction
Place Similarity Feature	This feature is used to capture the intuitions of the different places in terms of spelling, abbreviations and transliterations	Text Similarity is used to calculate the different places.
		Dice Co-efficient
		Jaro-Wrinlker Distance Co-efficient
		Double Metaphone Co-efficient
Place Type Similarity Feature (PTSF)	This feature vector represents the sense of proximity between the different types of places	Jaccard coefficient
		Dice Coefficient
		Semantic Similarity
Geospatial Similarity Feature (GSF)	This feature is used to represents the different locations of same places.	Normalized Distance Metrics
		Centroid Distance Metrics
		Normalized Distance Metrics
Semantic Relationship Feature (SRF).	The feature vector identifies the duplicate places which belongs to different places	Jaccard Coefficient
		Dice Co-efficient
Temporal Similarity Feature (TSF)	The feature vector identifies the similar temporal time intervals for the different places.	Temporal Time Duration
		Date Center Difference

3.4 ACTIVE SINGLE FEEDFORWARD LEARNING LAYERS:

In this it uses the principle of Extreme Learning Machines (ELM) for constructing the active learning models for the detection of data deduplication process. ELM was proposed by G.B.Huang[19], where the network can utilize the single hidden layer, higher speed and accuracy and preparing velocity, great speculation/exactness, and function with capabilities approximately[20,21].

In this system, the' L' neurons present in the hidden layer are needed to perform with an activation function that is differentiable like the sigmoid function, and the output layer is straight. in ELM, hidden layers do not require to be tunes mandatorily. The hidden layer compulsorily need not be tuned in ELM.

The loads of the hidden layer are appointed arbitrarily (counting the bias loads). It isn't the scenario are irrelevant for the hidden nodes, hence they do not require to be tuned and the hidden neurons parameters can be haphazardly fabricated in advance.

That is, before taking care of the training set data. For a single-hidden layer ELM, the system yield is given by eqn (1)

$$f_L(Y) = \sum_{i=1}^{L} \beta_i h1_i(Y) = h1(Y)\beta \quad (1)$$

Here Y is denotes the input

β denotes the output weight vector

$$\beta = [\beta_1, \beta_2, \ldots\ldots\ldots\ldots \beta_L]^T \quad (2)$$

H(Y) denotes the output hidden layer which is given by the following eqn

$$h1(Y) = [h1_1(Y), h1_2(Y), \ldots\ldots\ldots\ldots h1_L(Y)] \quad (3)$$

To determine Output vector O that can be defined as the target vector, the hidden layers are represented by eqn (4)

$$H = \begin{bmatrix} h1(Y_1) \\ h1(Y_2) \\ \vdots \\ h1(Y_N) \end{bmatrix} \quad (4)$$

The eqn represents the basic implementation of the ELM by using the minimal non -linear least square methods

$$\beta' = H^{*}O = H^{T}(HH^{T})^{-1}O \quad (5)$$

Here H∗ denotes the inverse of H known as Moore−Penrose generalized inverse.

Above eqn can also be represented as

$$\beta' = H^{T}(\frac{1}{C}HH^{T})^{-1}O \quad (6)$$

Hence the output function can be find by using the above eqn

$$f_L(Y) = h1(Y)\beta = h1(Y)\, H^{T}(\frac{1}{C}HH^{T})^{-1}O \quad (7)$$

ELM employs the kernel function for obtaining better accuracy for the higher performance. The benefit employing ELM are minimum training error and good approximation. The ELM employs the auto-tuning of the weight biases and non-zero activation functions. It finds its applications in classifying and predicting values. The description of its equations is found in [19],[22]. The complete architecture for the proposed ELM for predicting the duplication of data is shown in Figure 3

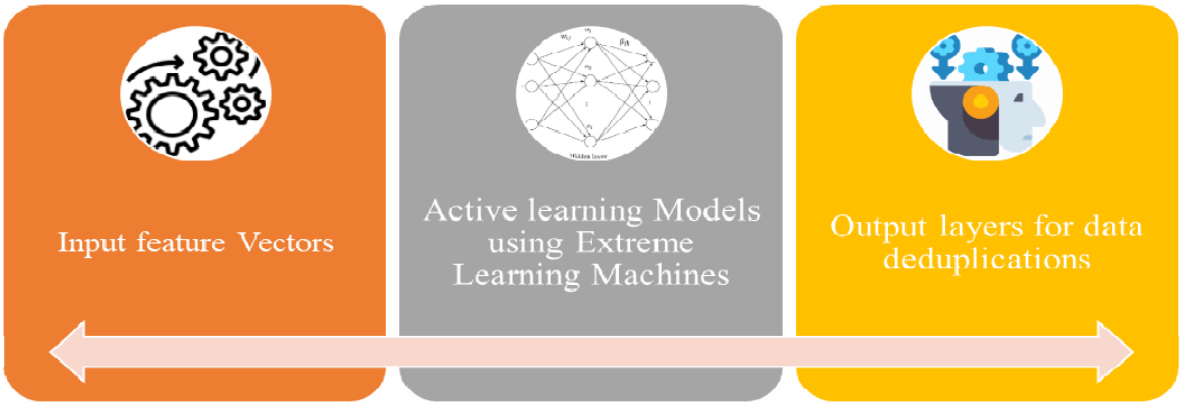

Figure 3 Proposed Framework for the Active Learning Models Based Data -Deduplication Identification process.

The parameters employed for training the proposed network is given in table III.

Table III Specifications Used for the Active Single Feed Forward Layers

Sl.no	Parameters used	Specifications
01	No of Neurons	100
02	Activation Function	Sigmoid
03	Input-Feature vectors Used	Multiple
04	No of Output Layers	02
05	Learning Error Rate	0.001

3.4 WORKING MECHANISM OF THE PROPOSED LEARNING MODELS:

After collection of the records, features extracted are used to train the proposed model to classify the duplicated data in the records. The features extracted are labelled for better identification of deduplication process. The complete work flow of the proposed algorithm is given in Algorithm-1

Sl.no	//Algorithm-1 -Pseudo Code for the Complete Working //
1	Collect the Gazzette Records (N) and store in the data bases
2	Extract the Features from the Records
3	While n=1 : L where L = maximum iteration
4	Train the Network with the Features
5	Determine the Value of fl using Equation(7)
6	If $f_l = F(t)$
7	//No duplication//
8	Else
9	// Duplication Found //
10	End
11	End

4. EXPERIMENTAL EVALUATION

The section-IV propounds about the performance metrics and results with comparative analysis between the proposed and existing algorithms.

4.1 PERFORMANCE METRICS:

As mentioned in Section 3, the dataset contains 1,257,00 gazetteer records for describing distinct places from all over the world. A total of 1,927 record pairs are manually labelled as duplicates. An analysis is made on the duplicate cases revealed at several situations acquired in the dataset that includes distinct place names with distinct spelling mistakes. Nearly 70% were taken as training data and 30% as testing the algorithm. The proposed learning models has been implemented using Python 3.8(Scikit learn) which runs on i5 CPU with 2TB HDD and 8 GB RAM. The following metrics were used for calculating the performance of the proposed algorithm is mentioned below

$$Accuracy = {TP + TN}/{TP + TN + FP + FN} \quad (8)$$

$$F - Score = \frac{2 * Precision * recall}{Precision + Recall} \quad (9)$$

$$\text{Precision} = {TP}/{TP + FP} \quad (10)$$

$$\text{Recall} = TP/(TP + FN) \quad (11)$$

$$Specificity = {TN}/{TN + FP} \quad (12)$$

where,

TN- True classified negative samples, TP- True positive samples

FP- False classified positives, FN- False classified negatives

The above metrics are used for the calculating the performance of the proposed algorithm. To prove the superiority of the proposed methodology, the research paper compares with the existing algorithms such as SVM and DT as mentioned in [23]. As mentioned in, all the feature vectors are taken as the input to all learning models. The accuracy in identifying the data deduplication using different learning models are shown in Figure 4

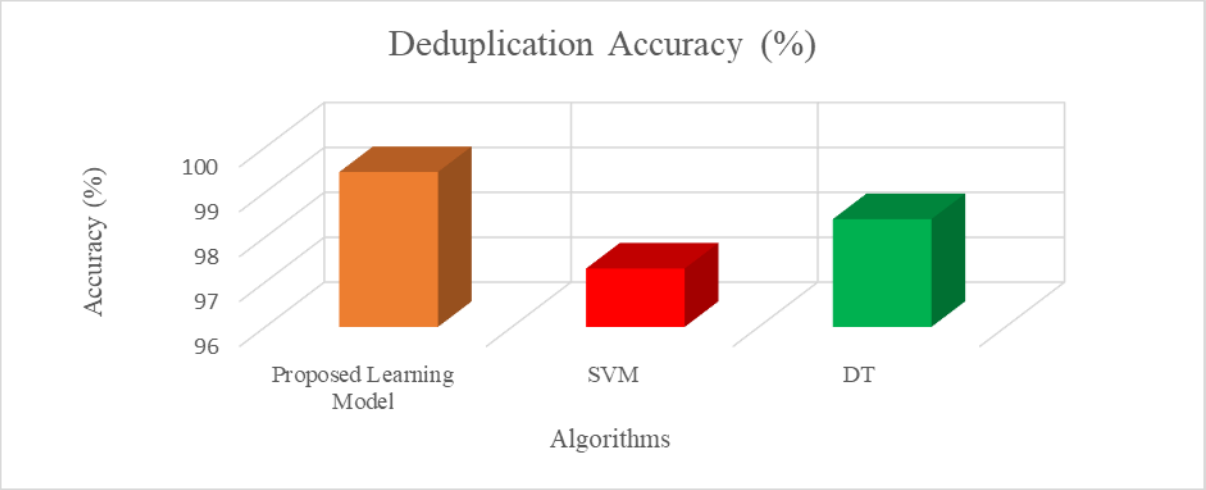

Figure 4 Deduplication Accuracy of the different learning models with the five categories of feature vectors.

From the Figure 4, it is found that the proposed learning model has exhibited the 99.45% deduplication accuracy whereas deduplication accuracy of SVM is 97.3% and decision tree is 98.3% respectively. The implementation of extreme learning machines 'properties has enhanced the deduplication accuracy by 2.15% than SVM and 1.15% than DT and proves proposed model will be suitable for identifying the data deduplication process. Moreover, the other performance metrics were also calculated and analysis which is presented in table IV.

Sl.no	Algorithm	Performance Metrics			
		Precision	Recall	F-Score	Specificity
01	Proposed Model	0.994	0.991	0.99	0.982
02	SVM	0.993	0.954	0.977	0.97
03	DT	0.976	0.97	0.98	0.971

From table IV, it is found that the performance of the proposed model has exhibited the edge of 1.5% and 2% over DT and SVM even for the larger datasets. From the above analysis, it is clear that proposed active single feedforward learning models has proved its efficiency in identifying the data deduplication process. Moreover, we have calculated the time complexity in classifying the data depilates and compared with the other existing algorithms. Figure 5 presents the time complexity analysis of different learning models in identifying the data deduplication process.

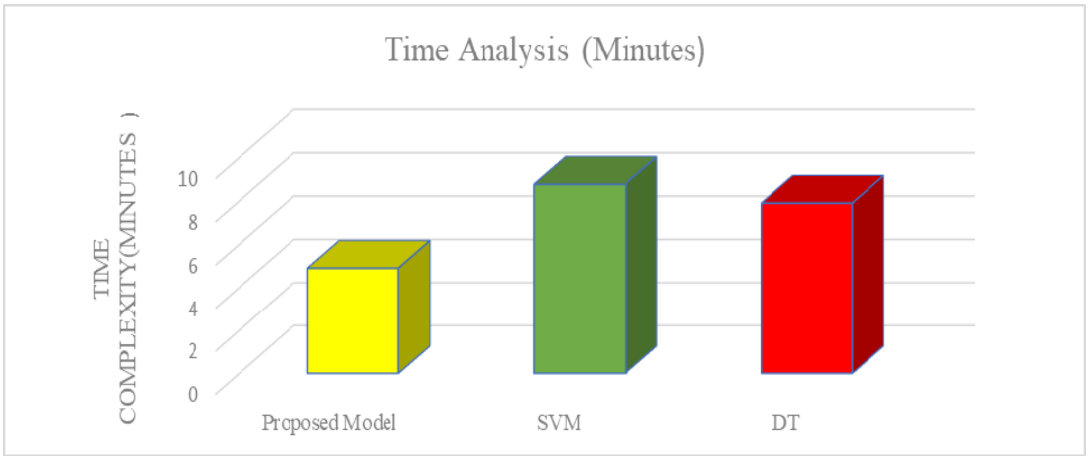

Figure 5 Time Complexity Analysis for the Different Learning Models for identifying the data deduplications.

From the Figure 5, it is clear that the integration of single feed forward layers in data deduplication process has less time to detecting when compared with the other existing algorithms.

5. CONCLUSION AND FUTURE SCOPE:

This paper presents the novel approach of implementing the active single feedforward layers in identifying the duplication in the data records. It is reported through the extraction of different features vectors were used as the input to the learning model. The feature vectors used for the identifying and classifying the data deduplication are 1) Place Similarity Feature (PSF)2) Place Type Similarity Feature (PTSF)3) Geospatial Similarity Feature (GSF) 4) Temporal Similarity Feature (TSF) and 5) Semantic Relationship Feature (SRF). The extensive analysis has been done by comparing the proposed model with the SVM and DT. From the analysis, it is found that proposed model has exhibited the deduplication accuracy of 99.45% and edge over the SVM with 97.4% and DT with 98.3% respectively. Though the proposed model has shown the promising results, but faces more challenges in future. This model has less time of detecting the data duplication which may suitable for cloud integration. But these models need more improvisation to handle the deduplication process for the huge datasets. Since the model doesn't incorporates any security schemes which may post the serious threats to the users. Hence the models with the most intelligent learning algorithms with an integration of high -end security schemes for data deduplication is mandatorily needed to ensure data integrity and an efficient usage of cloud spaces.

REFERENCES

[1] Naumann, F., Herschel, M., Ozsu, M.T.: An Introduction to Duplicate Detection. Morgan & Claypool Publishers (2010)
[2] Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity uncertainty and citation matching. In: Proceedings of the 7th Annual Conference on Neural Information Processing Systems (2003)
[3] Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man and Cybernetics 21(3) (1991) Duplicate Detection over Gazetteer Records 51\
[4] Samal, A., Seth, S., Cueto, K.: A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science 18 (2004)
[5] Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceedings of the 16th International Conference on Machine Learning (1999)
[6] Moguerza, J.M., Mu˜noz, A.: Support vector machines with applications. Statistical Science 21(3) (2006)
[7] Joachims, T.: Making large-scale SVM learning practical. In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods - Support Vector Learning. The MIT Press, Cambridge (1999)
[8] Witten, I.H., Frank, R.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
[9] M. Oh *et al*., "Design of Global Data Deduplication for a Scale-Out Distributed Storage System," *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, Vienna, Austria, 2018, pp. 1063-1073, doi: 10.1109/ICDCS.2018.00106.
[10] Zhu, R., Qin, Lh., Zhou, Jl. et al. Using multi-threads to hide deduplication I/O latency with low synchronization overhead. J. Cent. South Univ. 20, 1582–1591 (2013). https://doi.org/10.1007/s11771-013-1650-
[11] W. Xia et al., "FastCDC: A fast and efficient content-defined chunking approach for data deduplication," in Proc. USENIX Annu. Tech. Conf., 2016, pp. 101–114.
[12] D. Bhagwat, K. Eshghi, D. D. E. Long and M. Lillibridge, "Extreme Binning: Scalable, parallel deduplication for chunk-based file backup," *2009 IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems*, London, UK, 2009, pp. 1-9, doi: 10.1109/MASCOT.2009.5366623.
[13] Q. Zhang, M. F. Zhani, R. Boutaba and J. L. Hellerstein, "Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud," in *IEEE Transactions on Cloud Computing*, vol. 2, no. 1, pp. 14-28, Jan.-March 2014, doi: 10.1109/TCC.2014.2306427.
[14] W. Leesakul, P. Townend and J. Xu, "Dynamic Data Deduplication in Cloud Storage," *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, Oxford, UK, 2014, pp. 320-325, doi: 10.1109/SOSE.2014.46.
[15] Burramukku, Tirapathi & Rao, M.V.P.. (2017). Data deduplication in cloud storage using dynamic perfect hash functions. Journal of Advanced Research in Dynamical and Control Systems. 9. 2121-2132. 10.5013/IJSSST.a.19.04.08.
[16] S. Chavhan, P. Patil and G. Patle, "Implementation of Improved Inline Deduplication Scheme for Distributed Cloud Storage," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2020, pp. 1406-1410, doi: 10.1109/ICCES48766.2020.9137963.
[17] S. Chen, Y. Liang, Y. Chang, H. Wei and W. Shih, "Boosting the Profitability of NVRAM-based Storage Devices via the Concept of Dual-Chunking Data Deduplication," *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Beijing, China, 2020, pp. 512-517, doi: 10.1109/ASP-DAC47756.2020.9045622.
[18] P. Hamandawana, A. Khan, C. Lee, S. Park and Y. Kim, "Crocus: Enabling Computing Resource Orchestration for Inline Cluster-Wide Deduplication on Scalable Storage Systems," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 8, pp. 1740-1753, 1 Aug. 2020, doi: 10.1109/TPDS.2020.2972882.
[19] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," Neurocomputing, vol. 70, no. 1, pp. 489–501, 2006.
[20] Wang B, Huang S, Qiu J, et al. Parallel online sequential extreme learning machine based on MapReduce. Neurocomputing 2015; 149: 224-32.
[21] Lu S, Lu Z. A pathological brain detection system based on kernel-based ELM. Multimed Tool Appl 2016; 1–14.
[22] C. Lian, Z. Zeng, W. Yao, and H. Tang, "Displacement prediction of landslide based on psogsa-elm with mixed kernel," in Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference on. IEEE, 2013, pp. 52–57
[23] Bruno Martins," A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records", IEEE 2018 PP56-60

Authors Profile

Mr. N. Lakshmi Narayana, Research Scholar in Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram currently working as Assistant Professor at St. Ann's College of Engineering & Technology, Chirala. He is currently doing Research in the field of Cloud Computing, Network Security and Machine Learning.

Dr.B.Tirapathi Reddy, Associate Professor, Department of CSE,KLEF. obtained B. Tech from ANU, M. Tech from JNTUK, PhD from Acharya Nagarjuna University, Guntur. Dr. Reddy currently working as an Associate Professor in the Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram. Has 16 years of Academic & research experience, Published over 25 research Articles in National and International Journals. He is doing active research in the field of Data Science, Machine learning, Cloud Computing and Network security.