

Developed Modified Particle Swarm Optimization For Feature Selection On Learning Based Big Data In Cloud Computing

Thenmozhi L

Assistant professor, MGR College, Hosur, Tamilnadu.

Hosur, Tamilnadu

thenmozhilakshmanan@gmail.com

Dr.N.Chandrakala

HOD, SSM College of Arts and Science,

salem, Tamilnadu.

nchandrakala15@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10.

Abstract

The recent years, Machine Learning (ML) and Educational Data Mining (EDM) methods received significant attention for several aspects. Data Mining Methods can be employed for extracting meaningful data for the educational datasets. CMT (Cloud Mapping Tables), a result based mostly on utilization of information storage repository techniques for cloud-based Big - data analysis, can provide not only improved information management but also endurance with multidimensional information processing capacities. It could be utilized for a prediction of the current educational system, Student Academic Performance (SAP) has been used to appropriate corrective measures. . Data Storage is one of the best techniques to handle Big data. The information is frequently not even in the correct format to assist a company's decision-making process. In this paper, the issue addressed was applying a feature selection-based classification model for SAP. Initially, an adjusted Particle Swarm Optimization (MPSO) algorithm was applied effectively to select the set of features. For information categorization, the Multi-Layer Perceptron using Stochastic Gradient Descent (MLP-SGD) framework has been used. The suggested students ' performance approaches were combined with the Radial Basis Function (RBF) approach to reduce the miscategorized cases in the collected data, resulting in improved classification results. To verify the efficiency of the suggested, sets of data on the academic performance of students was obtained from secondary schools in Tanjore, Tamil Nadu, India. The modeled results demonstrated that the proposed significantly outperform the previous techniques.

Keywords: Modified Particle Swarm Optimization; Student Academic Performance; Data Mining; Machine Learning; Radial Basis Function; Cloud Computing; Big Data

1. Introduction

Big Data and Learning Analytics (LA) framework and plan powerful intercessions are basic to keeping up for scholastics limit. Information proficiency is a fundamental component of LA or the ability to make needed data adequately [1]. The instructive data study should be agreed with an instructively based game plan for fruitful and appropriate action result of investigation future test LA used the instructional models will, when all is said in done, instructive of current necessities reflect for instance, facing the instructive projects and the course passing and this complement dismisses the capriciousness of understudies associations with hazards isolating

understudies and teachers who are not masterminded to learn [2-4].

The Big Data period brings openings up in the learning investigation area for the arrangement of upgrading instructive material to students. In any case, the enormous information study gets huge inconveniences investigation, approval and prescient model turn of events [5]. In this paper, the creators present information-driven procedures for more noteworthy usage of learning study datasets, with the reason to improve the information on educators about a student's execution and furnish better personalization with advanced smart mentoring frameworks [6]. The proposed technique is unfurled in three phases. To begin with, the learning

investigation synopsis for introductory exploratory reasons for students' experience and their conduct in e-learning conditions [7].

Exactness instruction requires two similarly significant conditions: precise forecasts of scholastic execution dependent on early perceptions of the learning interaction and the accessibility of important instructive mediation choices [8]. The field of LA has made significant commitments to the acknowledgment of the main condition, particularly with regards to mixed learning and internet learning [9]. Expectation models that utilization information from institutional data frameworks and logs of learning the board frameworks have acquired a decent standing in foreseeing underperformance and dropout hazard. In any case, less advancement is made in settling the subsequent condition: applying LA-produced criticism to plan instructive mediations. LA gathers, studies, and reports huge information about students to streamline learning [10].

LA morals are an interdisciplinary field of study that tends to good, lawful, and social issues; thus, organizations are liable for executing systems that incorporate these points [11]. A considerable lot of the moral issues raised apply similar to instructive informational collections of any size. In any case, this study centers around large information that expand the scale and granularity of information assembled [12]. LA is giving new systems that are being applied to the planning and utilization of dashboards to help to instruct and learn. In any studies endeavor to see how teachers interface with an LA dashboard and how Self-directed Learning (SRL) exercises and feelings of educators happen and co-happen in the connection [13]. The current study researches ten educators' SRL exercises and epistemic feelings by dissecting the screen catch recordings and verbally processing information while they connect with an LA dashboard intended to help the online offbeat joint effort of numerous gatherings.

2. Literature Survey

Furnishing significant criticism with LA for understudies is an arising territory of exploration [14]. Past studies have regularly centered on course-level parts of supporting commitment with LA, however, understudies' points of view have gotten

restricted consideration [15]. The examined pre-administration instructors' necessities and assumptions for LA to help understudy commitment on the scholastic way level, which means noticing the continuum of study periods and scholarly years.

Neglecting to comprehend the viewpoints of teachers, and the requirements under which they work is a sign of numerous instructive innovation advancements' inability to accomplish utilization in genuine settings, and supported selection [16]. LA is no exemption, and there are progressively perceived strategy and execution challenges in advanced education for instructors to coordinate LA into their instruction [17]. Progressively LA has started using staff-and understudy confronting dashboards catching perceptions to introduce information to help understudy achievement and improve learning and education [18]. The utilization of LA is mind-boggling, diverse, and raises numerous issues for thought, including moral and legitimate difficulties, contending partner perspectives, and execution choices.

This study plans to recognize the theoretical design and the topical advancement in Learning Analytics and to expand spine/arising subjects in the field from 2011 to September 2019. To utilize various leveled grouping, key graphs, and organizational study to build the scholarly guide of the Learning Analytics people group and to imagine the topical scene in this field, utilizing co-word investigation [19]. Generally, an aggregate of 459 papers from the procedures of the Learning Analytics and Knowledge gathering and 168 articles distributed in the Journal of Learning Analytics, and the individual 3092 writers allotted catchphrases and 4051 machine-separated key-phrases, were remembered for the investigations.

Particle swarm optimization (PSO) would be a new method for optimization algorithms. The PSO was among the most prominent swarm-based models, & its search depended on models of social behaviors like an animal herd, fish schooling, & bird flocking, in which swarming collaborate to find food [20]. Because the PSO was simple to apply, it has grown in popularity in current years. For instance, a hybrid electric technique [21] that blends PSO & fuzzy logic-based linear programming. The PSO has been

used to solve real-world optimization issues with effectiveness.

In computer engineering, machine learning, systems analysis, & other related domains, optimization was crucial [22]. It should be the process of attempting to discover the optimal answer to an optimal solution in an acceptable amount of time [23]. An optimization issue aimed to find a value of a variable that meets the restrictions & maximizes or reduces the fitness value. The mathematical framework that could be utilized to tackle the issue was classified by the characteristics of fitness values to optimize. Evolutionary algorithms were a type of optimization procedure that stimulates the development of people through activities such as choice, variety, & reproducing. Health, or a person's competence in a given environment, drives this development.

3. Proposed framework

The entire mechanism of the proposed system performs on different levels like pre-processing, OD, FS, and classification. The MPSO-FS technique is used for selecting useful features and the MLP-SGD approach is used to classify the information generated from OD. A set of experiments is used to estimate the simulation outcome of the projected approach.

The CMT model utilizes CMT's cloud-based custom data center to assist academics and information scientists, while also presenting a technology that entails data analysis and extraction due to the efficient database management which can be accomplished quickly. With the expectation of still being utilized as a tool for the study and data analysts, we created the CMT with underlying goals in mind.

1. Big-data computing can be used to store information in a quite efficient manner.

2. Enables the use of customized information systems to analyze data, allowing for the classification of data into organized and unstructured categories.

3. Provide clients with just an interface that enables them to see content in various directions.

4. On this anonymized and responsible for describing, apply conventional data mining methods.

5. Increase the data access & information processing, i.e. versatility.

6. Take full advantage of cloud technologies at a lower cost.

7. Improve information visualization, costs, safety, querying assistance, usability, networking infiltration, and other Big - data processor concerns.

3.1 RBF based OD model

NN classifier goes under the classification of RBF non-direct classifier organization. At the point when ANN or NN is thought of, each MLP neuron acquires a weighted information esteem whole. Through a coefficient, each information esteem is duplicated with weighted worth, and the yields are added up to. Single MLP neurons are a straightforward direct classifier; notwithstanding, composite non-direct classifiers could be made by joining the neurons in an organization as demonstrated in Figure 1.

From the preparation dataset, the order of RBF Neurons do info's similitude measures towards likeness to tests. Out of the preparation set is an example that each RBF network is saved by a "model". Each neuron appraises its model among the Euclidean distance and contribution to arrange novel information. An info copies the models of class A contrasted with class B models and class A sorted it. RBF network design involves input vector and RBF neurons layer, and a yield layer with one hub or classification of information as demonstrated in Figure 2.

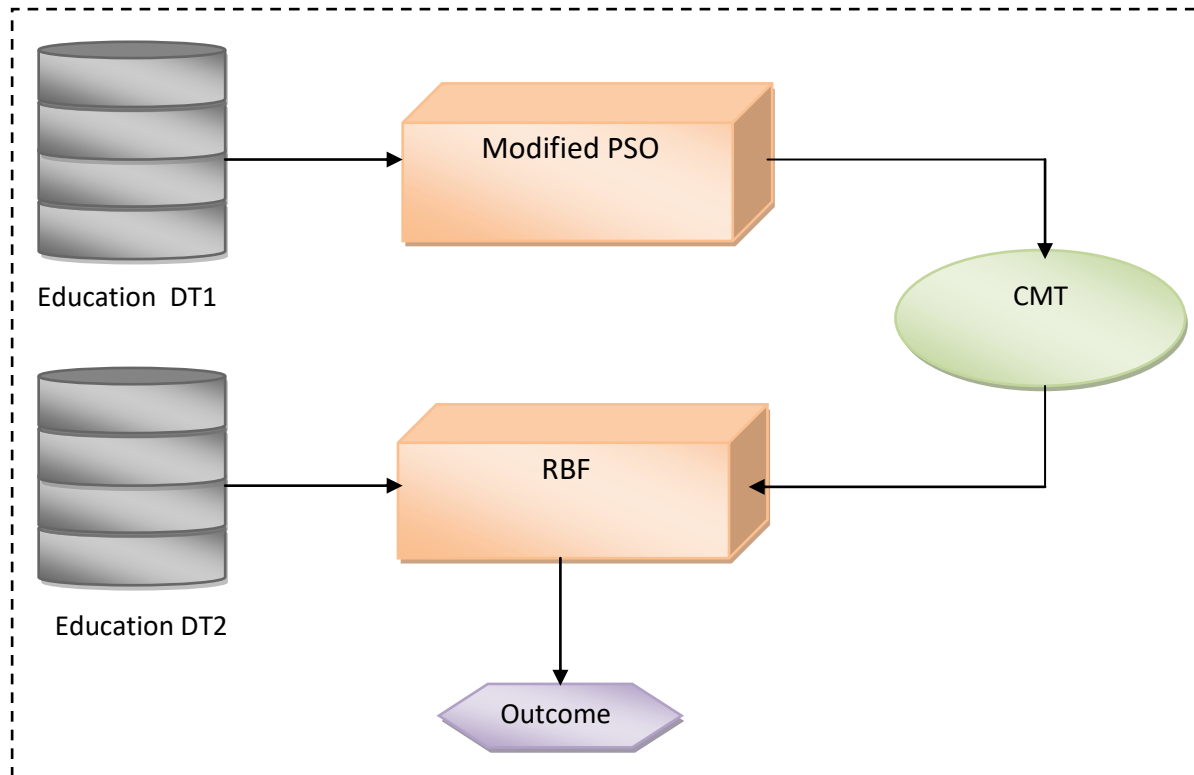


Figure 1 The overall process of proposed work

3.2 The Input Vector

The n-dimensional input pattern has been shown to all RBF cells & might be categorized. Each RBF cell in one vector stores the "prototype" vector from the training dataset. Each RBF neuron's input pattern was matched to its template, yielding a rate among 0 & 1, which serves as a classification model. When the input was equal to the concept, the RBF cell output was 1, & the reaction decreased as the gap between the model & input grows. The RBF neuron reaction structure would be a bell curve, as depicted in Figure 2 networked system. A neuron's response value is called the "activation" rate and the neuron's "center" was the prototype vector was located at a bell curve center.

Several nodes appear in the output of the system, one for each category that tries to classify the response. The classification decision is done by allocating the input to the classes that include a good score. From every RBF cell, the score was estimated through a weighted sum of activation rates and it was associated through the outcome node with every RBF

neuron. The neuron's activation is multiplied by all output nodes before calculating the response total. For various categories, each outcome node was estimated & comprises a quantity set. A good quantity will be offered for the outcome node towards RBF cells that depends on this category and to others, it produces negative weight.

3.3 RBF Neuron Activation Function

The same measures among its framework vector & input are estimated for each RBF cell. Input vectors were raising the similarity to the prototype gives an outcome towards 1. There exist various probable similarity functions choices, however; the highly popular function depends on Gaussian. With a 1D input, the Gaussian formula reveals in Equation 1. Let mean is represented through, the standard deviation is denoted through sigma & input by x. As seen in Figure 3, that results in a mean-centered bell-shaped curve.

$$\phi(x) = e^{-\beta \|x - \mu\|^2} \quad (1)$$

The mean of the Gaussian distribution is denoted as the prototype vector is represented as the bell curve

center.

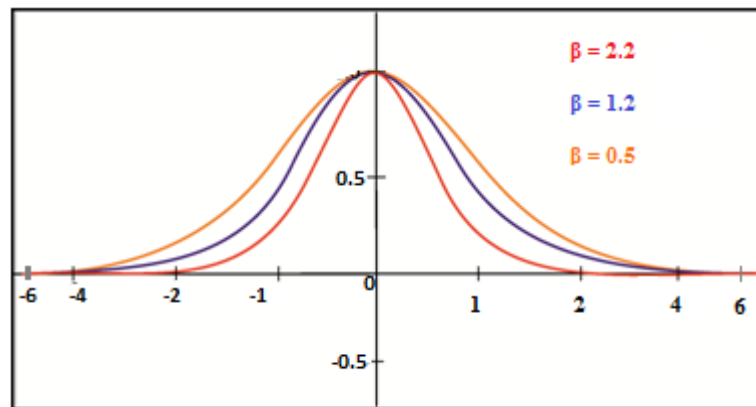


Figure 3 RBF Cell activation for Various values of beta

3.4 Feature selection procedure by MPSO-FS Algorithm

An MPSO-FS algorithm is designed for operating the FS process on the SAP technique. This model resembles the habit of training students in the classroom. The initial step is the teacher estimates the student with their corresponding knowledge and trained them accordingly. It cannot assure explicit factors and apply ordinary factors such as the number of inputs present in the population with higher rounds of operation. The proposed technique works based on a random searching model which has 2 levels: Teaching Phase and Learning Phase, as depicted in Figure 4.

Various classifiers are used in the FS process to identify the fitness value of everyone in the population. The ideal person was viewed as a trainer from the beginning, & he or she trains the remaining trainees by supposing them to be learners to improve

their knowledge depending on their abilities. Any pupil who achieves the highest accuracy was referred to as a teacher or professor. The teacher helps the learners to derive a solution along with the mean differing value. Prediction technique is applied for evaluating the population fitness in wrapper-based FS approaches.

Diverse classification models such as NB, SVM, k-Nearest Neighbors (KNN), DT, and Discriminant Analysis (DA) to identify the classification error rate and evaluate the precise value or fitness values of the system. A training factor is chosen randomly 1 and 2. Classification Accuracy is assumed as a fitness value to minimize the challenging issues of classification error rate. CA could be described as demonstrated in Equation (2).

$$CA = \frac{\text{Correctly Classified Instances}}{\text{Total Instances}} \quad (2)$$

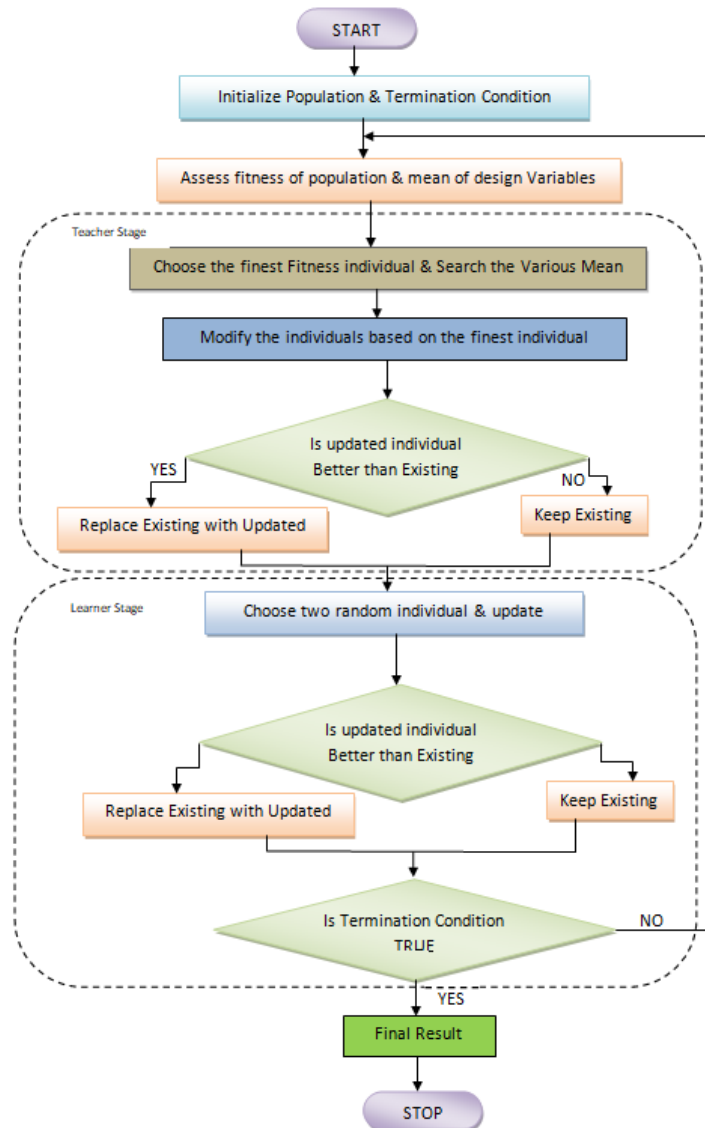


Figure 4 Flowchart of MPSO-FS

The second phase encompasses the improved skill of a learner, which is attained by communicating among the population itself. Every trainer presents in this solution space update the binary string using an alternate learner, which is chosen randomly from the population. The preferred individuals consist of optimal fitness values. An optimized FS provided the result with an individual lower error rate and greater accuracy that overcome the existing issues. Due to the rapid convergence of the classification error rate, this technique is executed with a greater number of productions that are restricted to 50 individuals. A

novelty dataset can be employed for training the classifier to obtain the maximum computation. It has undergone sampling with various population sizes and attained better results using 30 individuals. The main objective of the suggested system was:

- Decrease the number of characteristics to build a better classifier.
- Improve the accuracy of the training method.
- Decrease the time duration acquired for training ML methods along with choosing subsets.

When compared to the traditional FS technique, several experiments have been performed on the SAP

database to achieve optimal outcomes that have relevant results with ACO features. In the teacher phase, it exploits the present solution and in the learner phase, It finds a full-fledged search space for exploring the best results.

4. Performance Validation

The proposed method is implemented through MATLAB R2014a. The parameter setting follows: Crowd size is 50, Max_generation_size was 200 & Max_iteration_size was 200. For evaluating the

proposed SAP assessment model, a set of data is collected from a group of 1152 students through a questionnaire. The dataset is 42 in the total number of features present and the class count is 7 as The words "awesome," "wonderful," "outstanding," "excellent," "good," "standard," & "poor" have been used to describe the samples. The overall % attained under every class is 10.85%, 11.81%, 248%, 15.01%, 33.33%, 2.34% and 0.17% respectively.

Table 1 Confusion Matrix of MPSO Education dataset

	Very Good	Good	Awesome	Excellent	Outstanding	Average	Worst
Very Good	1	3	43	7	71	0	0
Good	0	9	44	6	77	0	0
Awesome	0	3	90	8	204	0	0
Excellent	1	7	56	2	107	0	0
Outstanding	0	1	78	4	301	0	0
Average	0	4	6	3	14	0	0
Worst	0	0	1	0	1	0	0

Table 2 Confusion Matrix of MPSO on cloud Education dataset

	Outstanding	Awesome	Excellent	Very Good	Good	Average	Worst
Outstanding	330	54	0	0	0	0	0
Awesome	118	187	0	0	0	0	0
Excellent	83	90	0	0	0	0	0
Very Good	64	61	0	0	0	0	0
Good	57	79	0	0	0	0	0
Average	13	14	0	0	0	0	0
Worst	0	1	0	0	0	0	0

Table 3 Confusion Matrix of MPSO on cloud Education Big dataset

	Outstanding	Awesome	Excellent	Very Good	Good	Average	Worst
Outstanding	259	1	0	0	0	0	0
Awesome	23	52	0	0	0	0	0
Excellent	3	5	0	0	0	0	0
Very Good	10	8	0	0	0	0	0
Good	10	11	0	0	0	0	0
Average	5	3	0	0	0	0	0
Worst	1	1	0	0	0	0	0

Tables 2, 3, and 4 show the confusion matrix offered by MPSO Education dataset, MPSO on cloud Education dataset, and MPSO on cloud Education

Big dataset respectively. The comparison of the FS findings acquired by the ACO-FS & MPSO-FS systems could be seen in Table 5.

Table 5 Matched analysis of FS with the proposed method

Methods	Best Cost	Selected Features
Proposed	0.034679	1,2,5,6,8,9,11,13,14,15,16,17,19,22,23,26,28,30,31,32,34,36,39,40
ACO-FS	0.052643	1,2,3,5,8,9,11,12,13,14,15,17,18,19,20,21,23,26,27,29,30,31,33,36,38,41

The FS results attained by the MPSO-FS and ACO-FS models are shown in Figures 5 and 6. The table values indicated that the ACO-FS model

achieved the maximum good rate of 0.052643 and a MPSO-FS system offered a minimum best cost of 0.034679.

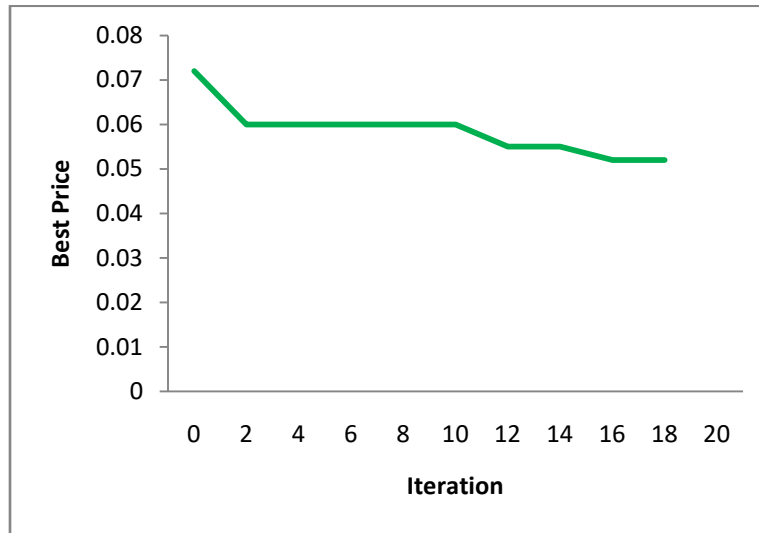


Figure 5 Best cost analysis of ACO-FS model

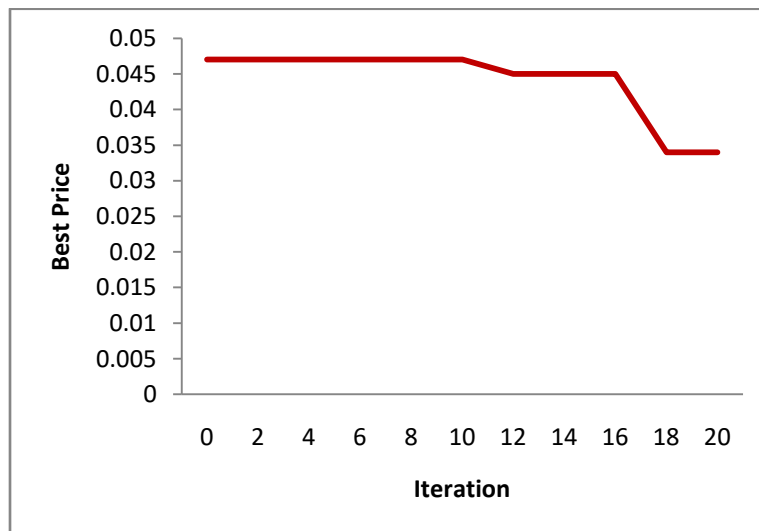


Figure 6 Best cost analysis of MPSO-FS model

5. Conclusions

Data Storage is one of the best techniques to handle Big data. The information is frequently not even in the correct format to assist a company's decision-making process. We can customize the education database system here just to handle the data but instead split it so that we can compute or analyze it as required. Hadoop's capacity to process information through processing multiple division equipment or infrastructures is similar to data partition. The rationale for the division is that we'll have a huge amount of data to analyze and don't have

the required capacity, we must divide the information and transfer it separately, otherwise it becomes garbage. FS process the number of features it aims to minimize and the search space size. This paper presents a new FS for SAP assessment that helps ML models. The proposed work operates on different levels, such as pre-processing, OD, FS, and classification. The MPSO-FS technique is executed for selecting useful features and the MLP-SGD approach is for classifying the information generated from OD. To evaluate the presented SAP assessment model, a set of data is collected from a group of 1152

students. The presented model offered maximum classification a higher precision of 98.73, recall of 74.85, an accuracy of 98.98, F score of 82.34 & kappa rate of 81.82. As a section of the next work, a presented work could be implemented in real-world scenarios.

References

- [1] Qi, Y., Ding, F., Xu, F., & Yang, J. (2020). Channel and feature selection for a motor imagery-based BCI system using multilevel particle swarm optimization. *Computational Intelligence and Neuroscience*, 2020.
- [2] Zhao, X., Turk, M., Li, W., Lien, K. C., & Wang, G. (2016). A multilevel image thresholding segmentation algorithm based on two-dimensional K-L divergence and modified particle swarm optimization. *Applied soft computing*, 48, 151-159.
- [3] Allam, M., & Nandhini, M. (2018). Optimal feature selection using binary teaching learning-based optimization algorithm. *Journal of King Saud University-Computer and Information Sciences*.
- [4] El-Kenawy, E. S., & Eid, M. (2020). Hybrid gray wolf and particle swarm optimization for feature selection. *Int. J. Innov. Comput. Inf. Control*, 16(3), 831-844. [5]
- [6] Thawkar, S. (2021). A hybrid model using teaching-learning-based optimization and Salp swarm algorithm for feature selection and classification in digital mammography. *Journal of Ambient Intelligence and Humanized Computing*, 12(9), 8793-8808.
- [7] Adedeji, O. T., Alade, O. M., & Oguntoye, J. P. (2021). Comparative Analysis of Feature Selection Techniques For Fingerprint Recognition Based on Artificial Bee Colony and Teaching Learning Based Optimization. *LAUTECH Journal of Computing and Informatics*, 2(1), 25-34.
- [8] Balamurugan, K., Uthayakumar, M., Sankar, S., Hareesh, U. S., & Warriar, K. G. K. (2020). Process optimization and exhibiting correlation in the exploitable variable of AWJM. *International Journal of Materials and Product Technology*, 61(1), 16-33.
- [9] Latchoumi, T. P., Reddy, M. S., & Balamurugan, K. (2020). Applied Machine Learning Predictive Analytics to SQL Injection Attack Detection and Prevention. *European Journal of Molecular & Clinical Medicine*, 7(02), 2020
- [10] Dash, C. S. K., Behera, A. K., Dehuri, S., & Cho, S. B. (2020). Building a novel classifier based on teaching-learning-based optimization and radial basis function neural networks for a non-imputed database with irrelevant features. *Applied Computing and Informatics*.
- [11] Arunkarthikeyan, K., & Balamurugan, K. (2020, July). Performance improvement of Cryo treated insert on turning studies of AISI 1018 steel using Multi-objective optimization. In *2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE)* (pp. 1-4). IEEE
- [12] Moradi, P., & Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing*, 43, 117-130.
- [13] Latchoumi, T. P., Kalusuraman, G., Banu, J. F., Yookesh, T. L., Ezhilarasi, T. P., & Balamurugan, K. (2021, November). Enhancement in manufacturing systems using Grey-Fuzzy and LK-SVM approach. In *2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)* (pp. 72-78). IEEE.
- [14] Ezhilarasi, T. P., Dilip, G., Latchoumi, T. P., & Balamurugan, K. (2020). UIP—A Smart Web Application to Manage Network Environments. In *Proceedings of the Third International Conference on Computational Intelligence and Informatics* (pp. 97-108). Springer, Singapore.
- [15] Şandru, E. D., & David, E. (2019, July). Unified feature selection and hyperparameter bayesian optimization for machine learning-based regression. In *2019 International Symposium on Signals, Circuits, and Systems (ISSCS)* (pp. 1-5). IEEE.
- [16] Latchoumi, T. P., Balamurugan, K., Dinesh, K., & Ezhilarasi, T. P. (2019). Particle swarm optimization approach for waterjet cavitation peening. *Measurement*, 141, 184-189
- [17] Balamurugan, K., Uthayakumar, M., Sankar, S., Hareesh, U. S., & Warriar, K. G. K. (2017). Mathematical modeling on multiple variables in machining LaPO₄/Y₂O₃ composite by abrasive waterjet. *International Journal of Machining and Machinability of Materials*, 19(5), 426-439.

- [18] Song, X. F., Zhang, Y., Gong, D. W., & Sun, X. Y. (2021). Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition*, 112, 107804.
- [19] Nemet, S., Ostojić, G., Kukolj, D., Stankovski, S., & Jovanovic, D. (2019). Feature Selection Using Combined Particle Swarm Optimization and Artificial Neural Network Approach. *Journal of Mechatronics, Automation and Identification Technology*, 4(1), 7-11.
- [20] Latchoumi, T. P., & Parthiban, L. (2022). Quasi oppositional dragonfly algorithm for load balancing in cloud computing environment. *Wireless Personal Communications*, 122(3), 2639-2656.
- [21] Banu, J. F., Muneeshwari, P., Raja, K., Suresh, S., Latchoumi, T. P., & Deepan, S. (2022, January). Ontology Based Image Retrieval by Utilizing Model Annotations and Content. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 300-305). IEEE.
- [22] Karnan, B., Kuppusamy, A., Latchoumi, T. P., Banerjee, A., Sinha, A., Biswas, A., & Subramanian, A. K. (2022). Multi-response Optimization of Turning Parameters for Cryogenically Treated and Tempered WC-Co Inserts. *Journal of The Institution of Engineers (India): Series D*, 1-12.
- [23] Ibrahim, R. A., Ewees, A. A., Oliva, D., Abd Elaziz, M., & Lu, S. (2019). Improved salp swarm algorithm based on particle swarm optimization for feature selection. *Journal of Ambient Intelligence and Humanized Computing*, 10(8), 3155-3169.

***Corresponding Author:**

Miss L.THENMOZHI:



I am L.THENMOZHI is a Research Scholar from the Department of Computer Science, Periyar University and working in MGR College, Hosur, krishnagiri District, Tamil Nadu, India. I am having 7 years of Teaching experience. My major area of research include Big Data, Computer Networks, Data Structure and Algorithms. I already published more than 5 publications in refereed journals.

Co-Author:

Dr.N.CHANDRAKALA :



Dr.N.CHANDRAKALA is HOD- PG & research Department of Computer science working in SSM College of Arts & Science, Komarapalayam, Namakkal District ,TamilNadu,India. She has 20 years of teaching experience. Her major areas of research include Computer Networks, Mobile computing and Big Data. She has more than 15 publications in refereed journals. She is acted as the reviewer of few journal.