ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20221119040501
Original scientific paper

MRMR-EHO-Based Feature Selection Algorithm for Regression Modelling

Sathishkumar V. E., Yongyun CHO*

Abstract: In the classical regression theory, a single function model is fit to a data set. In a complex and noisy domain, this process is too complex and/or not reliable. Piecewise regression models provide solutions to overcome these difficulties. The regression performance can be improved by proper feature selection. This paper proposes a feature selection technique for improving regression problems using the hybridization of filter and wrapper feature selection methods. It uses a hybrid framework of Elephant Herding Optimization (EHO) and minimum Redundancy and Maximum Relevance (mRMR). The mRMR-EHO is implemented to maximize the performance of individual regression algorithms and the results are provided in this research. In this paper, the effectiveness of CUBIST and mRMR-EHO feature selection using six fine grained data from small-sized data to big data is empirically demonstrated such as: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliances energy consumption dataset, f) Capital Bike share program data the results show a marginal increase in performance even to a very large scale. All 6 datasets were pre-processed well for building the models. The empirical results are based on the following algorithms: a) Generalized Linear Regression, b) K nearest neighbour, c) Random Forest, d) Support Vector Machine, e) Gradient Boosting Machine, f) CUBIST. Their performances are compared, and the best-performing model is selected. Ultimately, this paper puts forth that the mRMR-EHO-based feature selection with the rule-based CUBIST model for regression can be used as an effective tool for predictive data modelling in various domains.

Keywords: data mining; elephant herding optimization; feature selection; machine learning; MRMR

1 INTRODUCTION

The present decade is a dataset explosion age and deals with the exponential growth of data, which takes a lot of time and money if existent computers and algorithms are used. The modern era is challenged to handle big data in the global context. The size of the data bloats up both column-wise and row-wise in most of the applications. Several applications contribute to generating big data namely: healthcare, social media, transportation, Environmental monitors, Accounting systems, Logistics systems, and Personal devices. Due to the continuous generation of big data from several sources, problems will arise while accumulating the data in a centralized storage medium. And the technologies available to handle this big data need attention. The presence of noisy, irrelevant, and redundant data will affect the quality of the data. Technologies based on Artificial Intelligence, Machine Learning, and Deep Learning serve as efficient ways to handle this data crisis. Machine Learning is a simple and cost-efficient way to handle big data.

To use the generated big data for machine learning applications, the data should be preprocessed in such a way that a machine learning algorithm can effectively learn the data. While analyzing the data in high-dimensional spaces, several problems may occur. Bellman states this scenario as "The curse of Dimensionality" [1]. It gradually reduces the performance of the learning model, and it leads to overfitting. To overcome the problem with an increase in dimensions, one needs to reduce the count of features required for the model development.

Dimensionality Reduction techniques played an essential role in the reduction of the feature count and can be applied as a preprocessing technique to select optimal features. The primary purpose of dimensionality reduction is to identify the feature subset (dimensions) that are of relevance to the target class. From the literature, it is observed that there is a couple of methods that reduce the dimensions of the data and are namely: 1) Feature Selection (FS), 2) Feature Projection (FP).

The principle of feature selection from the perspective of increasing predictive accuracy might be a tool for improving classification accuracy or reducing the dimension without reducing classification accuracy. Global optimization, random search, as well as heuristic search are examples of popular algorithms. The main goal of feature selection is to assess a functioning dataset that has been chosen based on a collection of requirements. The outcome of the algorithm and the efficiency of a classification algorithm are directly determined by the evaluation. A good evaluation criterion states that the selected subset has a lot of knowledge and even less consistency. Identifying an optimum subset of features is a difficult and time-consuming process. Meta-heuristics have recently proven to be a powerful and accurate tool for resolving a variety of optimization problems (including machine learning, data mining, engineering design, and feature selection). The efficiency of the learning algorithm can be improved by properly balancing the discovery and mining. One choice for obtaining a good balance is to use an ensemble model, in which two or more algorithms are merged to boost each algorithm's efficiency, with the resultant ensemble approach. In the feature selection process, feature subset selection is pivoted on the features' nature. According to Yu and Liu [2], the features available in any dataset are categorized into four types, namely 1. Irrelevant features, 2. Weakly relevant but non-redundant features, 3. Redundant features, and 4. Strongly relevant features. Fig. 1 illustrates the types of features available.

According to Yu and Liu [2], strongly relevant features are required for an optimal feature subset selection and cannot be removed. Weakly relevant features are not always required but may be required in feature selection based on certain conditions. Irrelevant features are completely not at all required. Redundant features are part of weakly relevant features but are replaced by another set of features. An optimal feature subset always contains strong and non-redundant features. This optimal feature subset is also known as a Markov blanket of the given class label [3]. The stated blanket's principle discards every

redundant as well as irrelevant features from the original feature space [4].

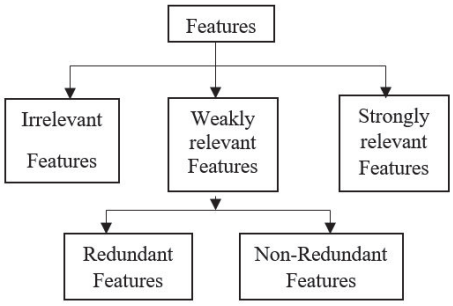

Figure 1 Classification of features

In general, the features need optimization. For the research on optimization, nature has been a significant source of strength. Various bio-inspired algorithms have been developed as a result. The architecture for such algorithms is largely the same. They begin with a set of randomly generated solutions. Then, using processes derived from nature, the collection is enhanced. In the Genetic Algorithm, each response is modeled after chromosomes in general, which are chosen based on their fitness value and undergo crossover and mutation with another chromosome. This emulates the natural selection, recombination, and mutation processes found in the evolution theory of the solution in Particle Swarm Optimization.

An algorithm changes the solutions unexpectedly during the discovery process, also known as diversification. This implies that the use of random components is at its highest level. The primary goal is to broaden the scope of the problem. A problem search space is *n*-dimensional, where *n* is the number of variables. The computational complexity may become certain (within the foam of discrete variables) or indefinite (within the foam of continuous variables) depending on the presence of the variables. In any case, to find the global optimum, an algorithm must search one of the most promising areas. This can be accomplished by modifying the variables at random. Experimentation should be accompanied by exploitation or intensification. Nature-inspired algorithms can regionally scan the provision region of a space search after identifying themselves. This will improve precision. To achieve this, reduce the size or rate of random variations in the solutions. Selecting the right ratio across exploration for a nature-inspired algorithm is challenging. Any type of feature selection can be used for the exploitation, and also the main complexity knowing when the local space search could have started. Some algorithms use efficient numbers and mechanisms to achieve a balance.

Despite belonging to the household of stochastic algorithms, according to the study, bio-inspired algorithms seem to be more convenient than conventional evolutionary algorithms (for example gradient-based). This is primarily due to these algorithms' superior local solution avoiding the problem independence has combined with traditional optimizations and search algorithms. The form of the search space distinguishes two types of optimization problems: linear and non-linear. Since there is no local solution during the first class, the best algorithm is a gradient-based strategy, which exceeds stochastic optimization algorithms substantially. It could be a significant count in local solutions in non-linear issues that should be ignored to find the global optimum. Although they step towards the negative gradient, gradient-based algorithms are prone to be trapped in local solutions. The stochastic components of nature-inspired algorithms help them escape certain local solutions in this case. Since they will not need a problem's gradient detail, they have a far broader range of uses than traditional algorithms. The EHO is an evolutionary algorithm.

This paper aims to discuss aspects of the optimal features selection using mRMR-EHO and best-performing regression model selection for the prediction of Strawberry Plants Nutrient water supply, Steel Industry Energy Consumption, Seoul Bike Sharing Demand, Seoul Bike Trip duration, Appliances energy consumption, and Capital Bike sharing system that should be considered when solving a regression problem in general, working along with the data mining based models to decide on the prediction problems of the considered datasets. By covering the questions including the requirement of the nutrient water supply, energy consumption, bike-sharing demand prediction, bike trip duration, Appliances energy prediction, and Capital Bike sharing system along with the selection of the model type through considering the influence of various feature influence strategy and taking into consideration the characteristics of the dataset at hand. Finally, it is to compare the results using Root Mean Squared Error values. The main contribution of this paper is summarized as follows:

1. To design a hybrid feature selection algorithm mRMR-EHO for improving the accuracy of the regression models.
2. Extracting optimal features from datasets with high dimensions.
3. To design a feature selection algorithm to reduce the search space.
4. To explore and study the related state-of-the-art approaches in the field of regression modeling using rule-based models and data analytics technologies.
5. To summarize and review relevant research to understand their proof of building predictive models and evaluation methods.
6. To collect data from several resources and Extract, Pre-process and Load the datasets.

1.1 Challenges and Motivation

Feature subsets selection or attribute subsets selection is useful in a lot of machine learning and data mining algorithms. The advantages of feature subset selection are manyfold. These advantages motivated us to deal with this problem. Different motivations and challenges are as follows:

Dimensional Curse: There are several features in the high-dimensional real-world data that adversely affect the efficiency of many of the data mining and machine learning algorithms. So, it is reasonable to keep only a limited number of input features and ignore others. It is a challenging task to decide which features to remove from the high-dimensional real-world data to improve the mining performance.

Reduce overfitting: There are situations in machine learning algorithms like a regression in which the classifier model fits precisely to the training data and there is a lot of difference in the predicted values and actual values when it is applied to test data. It leads to high training accuracy and low test accuracy; this is called the problem of overfitting. This issue occurs due to the existence of complexity and noisy features to make the dataset. It also does not help the mining algorithms to take any useful decisions. In the domain of bioinformatics for example when the purpose is to infer the most critical genes which can predict the class of patient whether he has cancer or not. The patient ID in the data is an irrelevant feature, if it is not removed from the data can give a different result. Therefore, feature selection is needed to reduce overfitting from the high-dimensional data.

Improve accuracy: The incredible amount of irrelevant feature characteristics in resistance to high increase the precision of different classifiers. Feature selection methods could solve this problem by selecting relevant features.

Improve efficiency: By applying a classification algorithm to the selected features can lead to reducing the computation time instead of applying it to all the features/attributes.

Extracting useful information: A dataset contains useful information that can help to analyze various processes. Such as cancer medical data assists the experts in disease diagnosis at an earlier phase. However, on the other hand, it is difficult to interpret and analyze such data as all the information is not relevant. Selecting useful genes from such data is a complex problem.

Handling high dimensional data: The main challenge in the data set is the availability of fewer sample numbers as compared to the number of features (genes). This collected data has a lot of class imbalance, which means that the number of a sample belonging to one class is more than the sample number belonging to another class. So, it is a thought-provoking task to develop a reliable feature selection algorithm that can handle any data.

2 RELATED WORKS

2.1 Works on Feature Selection

The literature is collected from different online repositories, library, and the focus is mainly on feature selection and regression. The related work presented here mainly focuses on dimension reduction using different feature selection algorithms for regression.

In the real world of big data, the available data contains a huge number of features. Each object is explained by hundreds of thousands or more attributes. So, analyzing my information from such data becomes a difficult task. It has become an important issue in almost all the spheres such as social media, e-commerce, healthcare, bioinformatics, transportation, computer vision, text mining, etc. The goal of dimensionality reduction is to keep only the relevant features from the data before it is given as input to any data mining or machine learning algorithm. The reduced data helps to give huge payoffs for decision-making by decreasing the learning time and by improving the learning performance. Moreover, pre-processing time decreases, and limited computational resources are required. High dimensional data may contain irrelevant and redundant features.

In the field of optimization, many researchers have adopted hybrid feature selection approaches in recent years. Hybrid algorithms outperformed traditional algorithms in a variety of practical and academic challenges [5]. The first metaheuristic-based feature selection methodology using an embedded technique was developed in 2004 [6]. Strategies based on local search were incorporated into the Genetic Algorithms (GA) based technique to manage the procedure of searching in this approach. Martin and Otto proposed a hybrid approach combining the simulated annealing with the Markov chain in [7], where the Markov chain is solely used for searching local optima. This algorithm was created to tackle the problem of travel salesman.

The reactive power issue [8] and the Symmetrical Traveling Salesman Problem have recently been solved using tabu search and hybrid simulated annealing algorithms [9]. In addition, in [10-13], Simulated annealing (SA) was hybridized utilizing a genetic algorithm. In comparison with other local or global search algorithms, the hybrid models performed better, according to this research. Numerous hybrid metaheuristic frameworks were developed with a high accuracy rate in the feature selection problem domain. In [14], Mafarja and Abdullah introduced a filter-based selection methodology that incorporated SA to improve the genetic algorithm's local search capacity. The results showed that the algorithm performed well on eight UCI datasets, resulting in a high number of chosen characteristics.

2.2 Works on Hybrid Feature Selection

Another wrapper feature selection hybrid technique that combines simulated annealing methodology with the crossover operator of GA was introduced in [15]. GA was combined with SA once more to create an embedded feature selection technique for classifying power disturbances in the Power Quality (PQ) issue and optimizing SVM parameters for the same [16, 17]. Olabiyisi et al. developed a hybrid GA-SA metaheuristic algorithm for feature extraction in the timetabling issue in 2012 [18], in which the GA selection phase was substituted by the SA selection process to evade stacking at local optima.

In a wrapper feature selection using the FUZZY ARTMAP NN classifier as an evaluator, GA was hybridized with Traveling Salesman [19]. Mafarja et al. [20] suggested two memetic filter feature selection methods. The fuzzy logic technique is used to regulate the primary feature in two local searching methodologies (record to record and huge deluge) that were later coupled with GA in these techniques. Moradi and Gholampour [21] devised a local search strategy to direct the search process in the Particle Swam Optimization (PSO) algorithm to find the smallest reduction constructed on correlation details. Furthermore, in [22], GA and PSO methodologies were combined with an SVM classifier termed GPSO in a hybridized technique for microarray data categorization. In the same way, [23] proposes a multi-phase PSO with a hybrid mutation operator. A novel GA-PSO combination

was presented in [24] to maximize the feature group for the Digital Mammogram database.

Two-hybrid wrapper parameter selection framework based on hybridization amongst Ant Colony Optimization (ACO) and GA [25, 26]. ACO algorithm was combined with Cuckoo Search (CS) algorithm in [27]. For the wrapper feature selection method, a hybrid model of the Harmony Search Algorithm (HSA) and a Stochastic Local Search (SLS) was proposed in [28]. For this purpose, Artificial Bee Colony (ABC) has been associated with a differential evolution algorithm (DE) [29]. Details about metaheuristics and feature selection can be obtained through survey papers in [30]. Tab. 1 summarizes the hybrid feature selection algorithms with their corresponding applications applied.

Table 1 Summary of hybrid feature selection algorithms

Reference	Algorithm	Application
[7]	Simulated Annealing + Markov Chain	Traveling salesman Problem
[9]	Tabu + Hybrid Simulated annealing algorithm	Symmetrics Traveling salesman Problem
[14]	SA + GA	Rough set reduction
[15]	SA + GA	Power Quality
[17]	GA − SA	Timetabling issue
[22]	GA + PSO	Microarray data classification
[24]	GA + PSO	Digital Mammogram database
[25]	ACO + GA	Text feature classification
[27]	ACO + CS	Digital Mammogram database
PROPOSED	mRMR + EHO	Smart farm, Energy Consumption, Transportation Application

New hybrid techniques can be proposed for feature selection apart from all the above-mentioned algorithms. The strategy in the optimization of a group called NoFree-Lunch (NFL) mentioned that all optimization problems cannot be solved by a single algorithm. It can be restated here as none of the heuristic wrapper feature selections could solve all feature selection-oriented problems. Alternatively, there is a space for improvements in the current techniques to provide better solutions. Hence another hybrid algorithm is proposed for feature selection. To contribute to hybrid feature selection methods, the mRMR-EHO-based hybrid method is proposed.

3 PROPOSED SYSTEM

3.1 Minimum Redundancy and Maximum Relevance (MRMR)

The feature vector is optimized using mRMR (minimal Redundancy and Maximum Relevance). The mRMR-obtained subset of features is both time-saving and accurate. Features having a strong correlation to the class (output) and a low correlation to other features are preferred in the mRMR's feature selection technique. Correlation with the class (relevance) may be determined for continuous characteristics using the *F*-statistic, and correlation between features can be determined using the Pearson correlation coefficient (redundancy). Then, a greedy search is used to choose features one-by-one in an effort to maximize the objective function, which is a function of the relevance and redundancy of each feature.

The objective function is often either a MID (Mutual Information Difference criterion) or a MIQ (Mutual Information Quotient criterion), where MID and MIQ stand for the difference and the quotient of relevance and redundancy, respectively. Some pre-processing procedures are needed to standardize temporal data into a single matrix before applying mRMR feature selection methodology. This might lead to a loss of potentially relevant information in time series (such as temporal order information).

3.2 Elephant Herding Optimization (EHO)

Because of their social nature, elephants form family groups consisting of mothers and their calves. A group of elephants living together is often led by a matriarch. Females prefer to reside with the family members, but males frequently choose to live elsewhere. They will begin to rely less and less on their family until eventually they leave home entirely. By researching the herding behavior of elephants, Elephant Herding Optimization (EHO) is proposed.

EHO comes with two unique operators: a clan updating operator and a separating operator. Whenever there is a change in the clan matriarch's status, the elephants are informed. Scholars and scientists are aware of EHO because of its acceptable performance. The elephants' herding behavior is modelled as two operators, and then idealized to create a generic global optimization strategy.

EHO takes into account the following hypotheses:
(1) Clans with fixed number of elephants comprise the elephant population.
(2) Every generation of elephants has a set of male elephants that decide to leave their family group, moving away from the rest of the herd and live solitarily.
(3) Each clan of elephants is headed by a matriarch.

3.3 MRMR-EHO

Elephant Herding Optimization (EHO) Algorithm is used for the feature subset selection process and the pre-process for attributes or parameters has been implemented by the mRMR technique. Before applying the feature subset selection module, the mRMR technique eliminates irrelevant and redundant features. The overall structure of the developed framework is depicted in Fig. 2. For a better solution, the proposed methodology embeds mRMR and EHO which are based on subset selection.

3.4 Dataset

For the process of feature selection: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliance's energy consumption dataset, f) Capital Bike share program data is used.

Strawberry Plants Nutrient water supply data include Temperature, Humidity, and CO_2. Steel Industry Energy Consumption data variables include lagging and important reactive power, the factor based on current power, emission of carbon dioxide, and load stages. Seoul Bike-sharing

demand data variables include weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information. Seoul Bike trip duration data variables include trip duration, trip distance, pickup, and drop off latitude and longitude, temperature, precipitation, wind speed, humidity, solar radiation, snowfall, ground temperature, and 1 hour average dust concentration. Weather data is collected from Korean Meteorological Administration. All four datasets were preprocessed well for building the models. Appliance's energy prediction dataset includes measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station, and recorded energy use of lighting fixtures. Capital Bike Share program data variables are the same as Seoul Bike Sharing demand data excluding the Functional days' attribute. Each of the datasets was independently applied for feature selection. The input for the algorithm must be a feature subset after pre-processing the whole feature subset.

3.5 MRMR and EHO-Based Feature Selection

The feature subset gathered using the dataset is larger in size but holds unwanted and non-essential features. These irrelevant parameters in the dataset affect the accuracy of regression and require a large duration for processing. To avoid this problem, there is a necessity for a significant selection of features on this feature subset to gather a significant feature set. So, a novel hybrid mRMR-EHO feature selection method is proposed for optimal feature selection in Fig. 2.

3.6 Ensemble-Based Regression

Six algorithms are used for the ensemble model of regression analysis: a) Generalized Linear Regression, b) K nearest neighbor, c) Random Forest, d) Support Vector Machine, e) Gradient Boosting Machine, f) CUBIST. After feature selection, each of the datasets is split into a training and testing dataset. The performance in the testing dataset is considered for analysis.

4 SCHEMATIC OUTLINE OF THE RESEARCH

All the datasets: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliance's energy prediction dataset, f) Capital Bike share program, are preprocessed and the final dataset is subjected to mRMA-EHO feature selection. The optimal features from mRMR-EHO are taken as the optimal feature subset for evaluation.

The feature subset is randomly split into training and testing sets: 75% for training and 25% for testing. Six models (GLM, KNN, RF, SVM, GBM, CUBIST) are trained with a training set utilizing 10 fold validation which is iterated thrice and their significant prediction accuracy is estimated using *RMSE* value. The overall procedure involved in this research is depicted in Fig. 3. The combined data set is split into training and testing sets using the train_test_split function. 75% of data is employed for training the models while the remaining 25% is used for testing.

Table 2 Training and testing set dimensions

Dataset	Training	Testing
Strawberry	164 and 13 variables	68 and 13 variables
Steel Industry Energy data	26,281 and 18 variables	8759 and 18 variables
Seoul bike sharing	6571 and 26 variables	2189 and 26 variables
Seoul bike trip	7200854 and 24 variables	2400285 and 24 variables
Appliance Energy Consumption	14803 and 38 variables	4932 and 38 variables
Capital Bike sharing system	13034 and 25 variables	4378 and 25 variables

The training and testing data dimensions for the 6 datasets: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliances energy prediction dataset, f) Capital Bike share program, are shown in Tab. 2.

For evaluating each of the regression models, the error values and the model fit need to be measured. Root mean squared error (*RMSE*) illustrates the sample deviation of the residue between the predicted and observed values. Residuals are a measure of the distance between the regression line and the data points. And *RMSE* is a measure of the residual's spread. In brief, it gives the concentration of the data that is around the line of best fit. Equation for computing *RMSE* is given in Eq. (1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \tag{1}$$

5 RESULTS AND DISCUSSION

All the results attained from the proposed model are depicted in this section. In this section, the experimental results for datasets: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliance's energy prediction dataset, f) Capital Bike share program data using the prediction algorithms GLM, KNN, RF, SVM, GBM, and CUBIST were presented.

Table 3 Datasets description

Dataset	Instance	Attributes	Type of Attributes
Strawberry [31]	232	13	Real
Steel Industry Energy data [32]	35040	18	Categorical, Integer
Seoul bike sharing [33]	8760	26	Categorical, Integer
Seoul bike trip [34]	9601139	60	Categorical, Integer
Appliance Energy Consumption [35]	19735	29	Categorical, Integer
Capital Bike sharing system [36]	17389	25	Categorical, Integer

All the models are trained using their training set for six datasets: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliances energy prediction dataset, f) Capital Bike share program data, g) with their best hyper-parameters

chosen from the Grid Search Cross-validation results. The models are evaluated using the testing set with RMSE values. The details of the datasets used are shown in Tab. 3. Algorithms producing fewer RMSE values are considered as the best-performance algorithm.

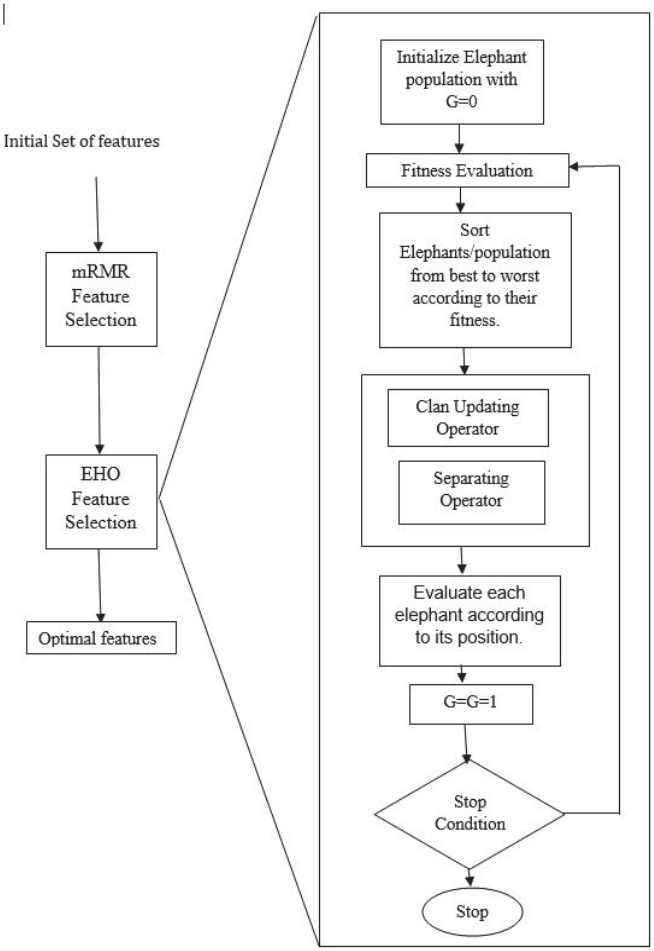

Figure 2 MRMR-EHO

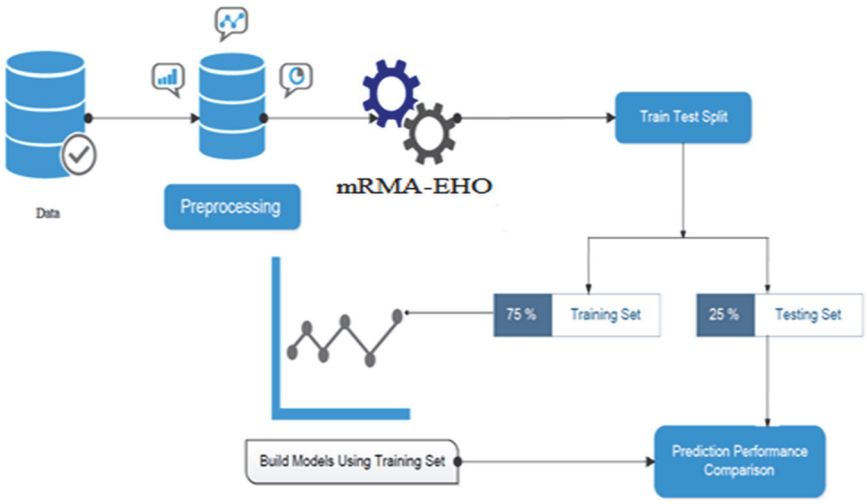

Figure 3 Schematic outline of the research

Table 4 RMSE values without feature selection

Dataset	GLM	KNN	RF	SVM	GBM	CUBIST
Strawberry	60.97	59.20	56.64	56.22	57.59	51.36
Steel Industry Energy data	4.61	3.30	1.12	1.97	0.53	0.24
Seoul bike sharing	424.61	299.88	243.98	242.89	172.73	139.64
Seoul bike trip	16.48	13.93	6.25	15.58	12.58	6.05
Appliance Energy Consumption	93.18	77.65	68.48	70.74	66.65	64.43
Capital Bike sharing system	432.45	298.65	253.54	313.43	243.23	214.32

Table 5 Attributes selected using MRMR and MRMR-EHO

Dataset	Total Attributes	mRMR Attributes	mRMR-EHO
Strawberry [31]	13	10	8
Steel Industry Energy data [32]	18	15	14
Seoul bike sharing [33]	26	23	20
Seoul bike trip [34]	60	52	48
Appliance Energy Consumption [35]	38	31	28
Capital Bike sharing system [36]	25	22	20

Table 6 RMSE Values with mRMR features

Dataset	Total Attributes	mRMR Attributes	GLM	KNN	RF	SVM	GBM	CUBIST
Strawberry	13	8	58.34	52.87	52.62	50.63	55.38	48.56
Steel Industry Energy data	18	14	3.78	2.94	1.05	1.73	0.49	0.23
Seoul bike sharing	26	20	410.54	284.93	239.38	237.83	169.45	135.79
Seoul bike trip	60	48	15.73	12.96	6.01	14.56	11.09	5.97
Appliance Energy Consumption	38	28	90.68	72.21	62.56	67.71	62.88	59.77
Capital Bike sharing system	25	20	422.99	281.43	246.67	307.32	232.11	202.28

Table 7 RMSE Values with MRMR-EHO features

Dataset	Total Attributes	mRMR-EHO	GLM	KNN	RF	SVM	GBM	CUBIST
Strawberry	13	8	53.71	48.32	48.64	46.22	51.59	42.36
Steel Industry Energy data	18	14	2.83	2.74	0.98	1.53	0.45	0.21
Seoul bike sharing	26	20	399.38	267.54	232.81	222.43	161.52	129.64
Seoul bike trip	60	48	13.67	10.87	5.43	12.26	10.45	4.78
Appliance Energy Consumption	38	28	86.45	65.23	56.67	62.54	57.87	51.34
Capital Bike sharing system	25	20	405.58	268.75	233.54	298.52	219.98	189.43

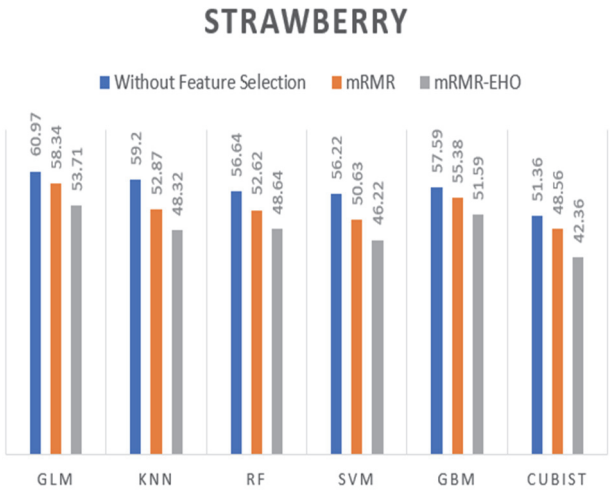

Figure 4 Strawberry data RMSE values comparison

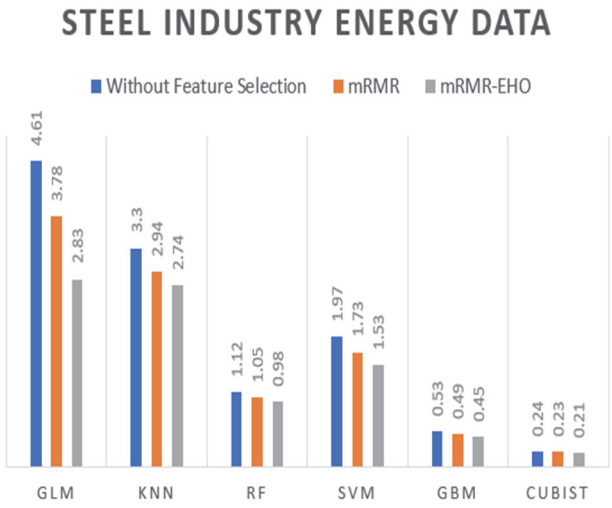

Figure 5 Steel industry energy data RMSE values comparison

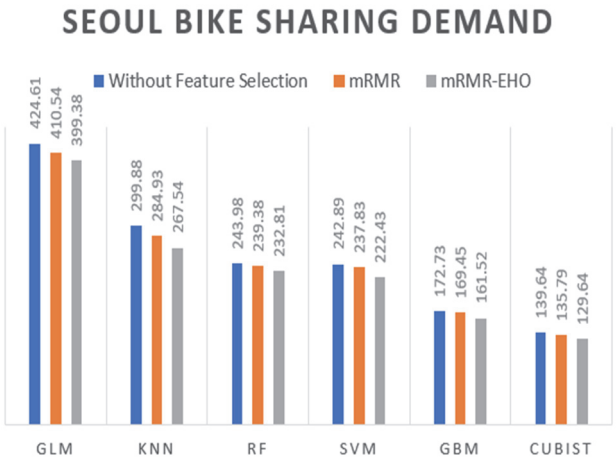

Figure 6 Seoul bike sharing demand data RMSE values comparison

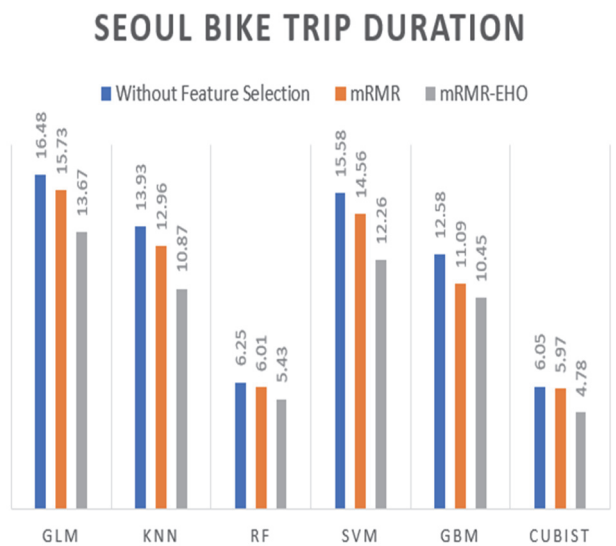

Figure 7 Seoul bike trip duration data RMSE values comparison

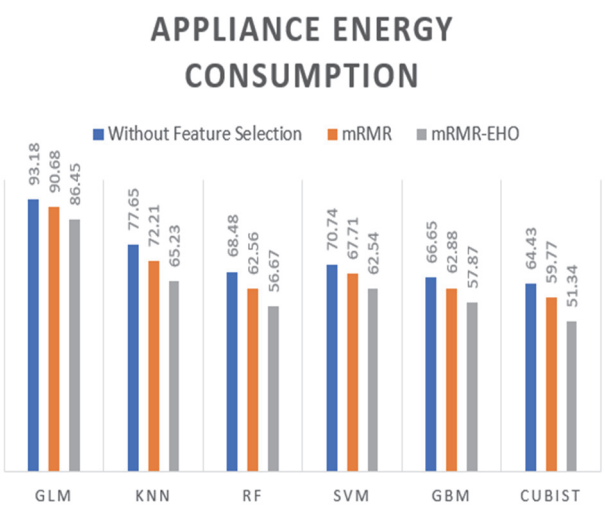

Figure 8 Appliance Energy Consumption data RMSE values comparison

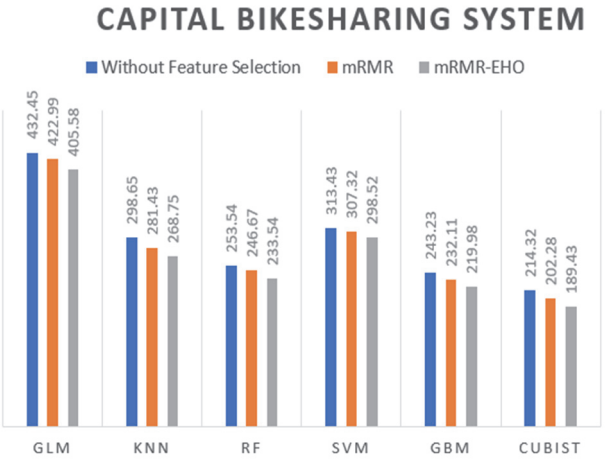

Figure 9 Capital bike sharing data RMSE values comparison

RMSE values of the datasets without feature selection methods are shown in Tab. 4. As can be seen from the table, CUBIST algorithm performance is better than GLM, KNN, RF, SVM, and GBM and leaves a conclusion that CUBIST algorithm performance is better than conventional algorithms in each of the domains considered in this study.

The number of attributes selected using mRMR and mRMR-EHO is summarized in Tab. 5. It can be observed from Tab. 5 that the number of features selected by mRMR is less than the Total attributes. This shows redundant features are removed by mRMR. This in turn enhances the speed of training. After filtering features using mRMR, the dataset is further subjected to mRMR-EHO feature selection. The number of features is further reduced by mRMR-EHO (Tab. 5). The resulting number of features from the mRMR-EHO feature selection algorithm is shown in Tab. 5.

When compared to Tab. 4, RMSE without feature selection with Tab. 6, RMSE value with mRMR features, the RMSE values for all the algorithms considered in this study are improved considerably. Even now in Tab. 8 CUBIST algorithm with mRMR features outperforms other algorithms in consideration. Results from GLM show that all the datasets are not linearly distributed over independent variables.

Tab. 7 presents RMSE values with mRMR-EHO features. As can be seen from the table, RMSE values are further improved compared to Tab. 5 and Tab. 6. This proves that feature selection using mRMR-EHO hybridization is best compared with mRMR feature selection and without feature selection. The feature selection using mRMR-EHO considerably improves the prediction performance of all the regression models.

Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9 show the comparison of the RMSE value for Strawberry Plants Nutrient water supply data, Steel Industry Energy Consumption data, Seoul Bike Sharing Demand data, Seoul Bike Trip duration data, Appliances energy prediction dataset data, Capital Bike share program data respectively. As can be seen from the figures feature selection using mRMR-EHO improves the regression performances for all the algorithms. Out of the conventional algorithms considered in this research, the CUBIST model outperforms all the models.

The results show that the performance of mRMR-EHO feature selection with the CUBIST algorithm is best compared to other traditional algorithms is considered domain and can be used as an effective tool in predictive analytics.

6 CONCLUSION

This study focused on predicting the dependent variables in four distinct domains using corresponding datasets: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliances energy prediction dataset, f) Capital Bike share program data. The value to be predicted in each dataset known as the dependent variable is predicted using Data Mining algorithms (GLM, KNN, RF, SVM, GBM, and CUBIST) with optimal features from the proposed hybrid feature selection algorithm mRMR-EHO. A hybrid of mRMR-EHO is proposed to optimize the feature selection process. The results show that mRMR and EHO-based feature selection improves the RMSE values in all the datasets. This shows that the mRMR and EHO-based feature selection can be used as an effective feature selection tool. Additionally, a rule-based model CUBIST is proposed in this research for the considered domains. The CUBIST model outperforms all the conventional algorithms in each of the domains. A comparison with the traditional algorithm is done to prove the efficiency of the proposed feature selection algorithm mRMR-EHO and CUBIST model. The results show that the mRMR-EHO-CUBIST algorithm improves the RMSE values compared to traditional models like GLM, KNN, RF, SVM, and GBM models in four datasets. This research proves that the mRMR-EHO feature selection method with the CUBIST framework can be utilized as a significant model in various domains like water consumption prediction, energy consumption prediction, bike-sharing demand prediction, and trip duration prediction. This model finds a significant association among the parameters.

This work illustrates the present literature in various ways. This study depicts the work based on the empirical framework of water consumption prediction, energy consumption prediction, bike-sharing demand prediction, and trip duration prediction with a novel hybrid feature

selection technique. The mRMR-EHO is based on the strategy of hybrid feature selection, an advanced framework of significant approach for improving the significance of feature selection and thereby improving the regression values.

Limitations of this study: This research is based on selecting optimal features using Minimum Redundancy and Maximum Relevance (mRMR)-Elephant Herding optimization (EHO) hybrid and fitting regression models for: a) Strawberry Plants Nutrient water supply, b) Steel Industry Energy Consumption, c) Seoul Bike Sharing Demand, d) Seoul Bike Trip duration, e) Appliances energy consumption, f) Capital bike-sharing system data and aims to build an optimum predictive model.

For Strawberry Plants Nutrient water supply prediction, we have not included nutrient water concentration details and external greenhouse weather information.

For Steel industry energy consumption prediction, we have not included weather data since the industry is a small-scale industry and situated in open space and so weather-related information has no correlation with energy consumption. For the appliances' energy consumption dataset, weather information is included but the occupancy details are not specified.

It does not include information on each customer's usage habits or the activities of each docking station when predicting Seoul bike-sharing demand and the capital bike-sharing system.

For Seoul bike trip duration prediction, it excludes the usage pattern of each customer, or each docking site activity information. Apart from the historical data on the trip duration of bikes used by the customers, this study includes trip distance and weather information. Other factors like traffic, pollution, government policy and so on, are excluded.

This study, which is based on data gathered while authoring this work, aims to train regression models with the best hyper-parameters from the proposed mRMR-EHO. Additionally, the usage of models is only permitted for a period of three to four months. Additionally, it is necessary to update the model over time as new data is added.

Future work, Water Supply prediction: in Greenhouse: Investigating the greenhouse's harvesting processes in the first stage. The second stage is to replace the greenhouse with historical data and the model can predict water demand so that the water consumption can be optimized. This will help the greenhouse to become a smart system with intelligent decision-making strategies.

Energy Consumption: Future studies may examine the energy consumption of specific pieces of equipment while taking into account how long they are used, and it may create software, like an alert system, that is activated when an unusual energy consumption trend is observed.

Bike-sharing demand: Predicting station level or dock level bike demands can help bike managers to provide uninterrupted bike supply at each dock. Employing clustering to group docks having the same demand patterns and applying machine learning algorithms based on each cluster could be a strategy to predict station-level demand. Addressing rebalancing issues in bike docks also seems to be an interesting problem.

Transportation Systems, Trip duration: Since the trip duration dataset has several entries, employing deep learning methods can increase prediction accuracy. Additional information about the trip duration such as traffic information, number of signals, etc. can be considered. By detecting trip duration in real-time, an analysis can be done to detect traffic on roadways and can be used to solve traffic problems while taking a route.

Acknowledgments

This paper is based on the Ph.D. thesis of Sathishkumar V E. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01489) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). Interdisciplinary Program in IT-Bio Convergence System (BK21 Plus), Sunchon National University, 255, Jungang-ro, Suncheon-si, Jeollanam-do 57922, Republic of Korea. This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea funded by the Ministry of Education (MOE) (2021RIS-002).

7 REFERENCES

[1] Powell, W. B. (2009). Approximate Dynamic Programming: Solving the curses of dimensionality. *IIE Transactions*, *41*(2), 168-169. https://doi.org/10.1080/07408170802189500

[2] Yu, L. & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, *5*, 1205-1224.

[3] Chun, Y. H. & Cho, M. K. (2022). An empirical study of intelligent security analysis methods utilizing big data. *Journal of Logistics, Informatics and Service Science*, *9*(1), 26-35. https://doi.org/10.14704/WEB/V19I1/WEB19311

[4] Lee, J. M., Jung, I. H., & Hwang, K. (2022). Classification of beef by using artificial intelligence. *Journal of Logistics, Informatics and Service Science*, *9*(1), 1-10. https://doi.org/10.14704/WEB/V19I1/WEB19308

[5] Kundu, T. & Garg, H. (2022). A hybrid ITLHHO algorithm for numerical and engineering optimization problems. *International Journal of Intelligent Systems*, *37*(7), 3900-3980. https://doi.org/10.1002/int.22707

[6] Oh, I. S., Lee, J. S., & Moon, B. R. (2004). Hybrid genetic algorithms for feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, *26* (11), 1424-1437. https://doi.org/10.1109/TPAMI.2004.105

[7] Martin, O. C. & Otto, S. W. (1996). Combining simulated annealing with local search heuristics. *Ann. Oper. Res.*, *63*(1), 57-75. https://doi.org/10.1007/BF02601639

[8] Lenin, K., Reddy, B. R., & Suryakalavathi, M. (2016). Hybrid Tabu search-simulated anneal- ing method to solve optimal reactive power problem. *Int. Electr. Power Energy Syst.*, *82*, 87-91. https://doi.org/10.1016/j.ijepes.2016.03.007

[9] Lin, Y., Bian, Z., & Liu, X. (2016). Developing a dynamic neighbourhood structure for an adaptive hybrid simulated annealing–tabu search algorithm to solve the symmetrical traveling salesman problem. *Appl. Soft Comput.*, *49*, 937-952. https://doi.org/10.1016/j.asoc.2016.08.036

[10] Vasant, P. (2010). Hybrid simulated annealing and genetic algorithms for industrial production management problems. *Int. J. Comput. Methods*, *7*(02), 279-297. https://doi.org/10.1142/S0219876210002209

[11] Li, Z. & Schonfeld, P. (2015). Hybrid simulated annealing and genetic algorithm for optimizing arterial signal timings

under oversaturated traffic conditions. *J. Adv. Transp*., *49*(1), 153-170. https://doi.org/10.1002/atr.1274

[12] Li, Y., Hao. G., Lin, W., & Jing, F. (2013). A hybrid genetic-simulated annealing algorithm for the location-inventory-routing problem considering returns under E-supply chain environment. *Sci. World J*, 1-11. https://doi.org/10.1155/2013/125893

[13] Junghans, L. & Darde, N. (2015). Hybrid single objective genetic algorithm coupled with the simulated annealing optimization method for building optimization. *Energy Build.*, *86*, 651-662. https://doi.org/10.1016/j.enbuild.2014.10.039

[14] Mafarja, M. & Abdullah, S. (2013). Investigating memetic algorithm in solving rough set attribute reduction. *Int. J. Comput. Appl. Technol.*, *48*(3), 195-202. https://doi.org/10.1504/IJCAT.2013.056915

[15] Reza, A., Boshra, P., Narges, N., Maryam, K., & Fahimeh, B. (2010). A hybrid GA and SA algorithms for feature selection in recognition of hand-printed Farsi characters. *Proceeding of the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*. https://doi.org/10.1109/ICICISYS.2010.5658728

[16] Wu, J. & Lu, Z. (2012). A novel hybrid genetic algorithm and simulated annealing for feature selection and kernel optimization in support vector regression. *Proceeding of the 2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI).* https://doi.org/10.1109/ICACI.2012.6463321

[17] Manimala, K., Selvi, K., & Ahila, R. (2011). Hybrid soft computing techniques for feature selection and parameter optimization in power quality data mining. *Appl. Soft Comput*., *11*(8), 5485-5497. https://doi.org/10.1016/j.asoc.2011.05.010

[18] Olabiyisi Stephen, O., et al. (2012). Hybrid metaheuristic feature extraction technique for solving timetabling problem. *Int. J. Sci. Eng. Res*., *3*(8).

[19] Tang, W. C. (2007). Feature Selection For The Fuzzy Artmap Neural Network Using A Hybrid Genetic Algorithm And Tabu Search, USM.

[20] Majdi, M., Abdullah, S., & Jaddi, N. S. (2015). Fuzzy Population-based meta-heuristic approaches for attribute reduction in rough set theory. *World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng*., *9*(12), 2289-2297.

[21] Moradi, P. & Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Appl. Soft Comput*., *43*, 117-130. https://doi.org/10.1016/j.asoc.2016.01.044

[22] Talbi, E. G., Jourdan, L., Garcia-Nieto, J., & Alba, E. (2008). Comparison of population based metaheuristics for feature selection: application to microarray data classification. *Proceeding of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, IEEE.* https://doi.org/10.1109/AICCSA.2008.4493515

[23] Yong, Z., Dun-wei, G., & Wan-qiu, Z. (2016). Feature selection of unreliable data using an improved multi-objective PSO algorithm. *Neurocomputing*, *171*, 1281-1290. https://doi.org/10.1016/j.neucom.2015.07.057

[24] Jona, J. & Nagaveni, N. (2012). A hybrid swarm optimization approach for feature set reduction in digital mammograms. *WSEAS Trans. Inf. Sci. Appl*., *9*, 340-349.

[25] Basiri, M. E. & Nemati, S. (2009). A novel hybrid ACO–GA algorithm for text feature selection. *Proceeding of the 2009 IEEE Congress on Evolutionary Computation*. https://doi.org/10.1109/CEC.2009.4983263

[26] R. Babatunde, S. Olabiyisi, E. Omidiora. (2014) Feature dimensionality reduction using a dual level metaheuristic algorithm. *International Journal of Applied Information Systems (IJAIS)*, *7*(1). https://doi.org/10.5120/ijais14-451134

[27] Jona, J. & Nagaveni, N. (2014). Ant-cuckoo colony optimization for feature selection in digital mammogram. *Pakistan J. Biol. Sci*., *17*(2). https://doi.org/10.3923/pjbs.2014.266.271

[28] Nekkaa, M. & Boughaci, D. (2016). Hybrid harmony search combined with stochastic local search for feature selection. *Neural Process. Lett*., *44*(1), 199-220. https://doi.org/10.1007/s11063-015-9450-5

[29] Zorarpacı, E. & Özel, S. A. (2016). A hybrid approach of differential evolution and artificial bee colony for feature selection. *Expert Syst. Appl*., *62*, 91-103. https://doi.org/10.1016/j.eswa.2016.06.004

[30] Boussaïd, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization, metaheuristics. *Inf. Sci*., *237*, 82-117. https://doi.org/10.1016/j.ins.2013.02.041

[31] Sathishkumar, V. E., Lee, M. B., Lim, J. H., Shin, C. S., Park, C. W. & Cho, Y. Y. (2019). Predicting Daily Nutrient Water Consumption by Strawberry Plants in a Greenhouse Environment. *Proceedings of the Korea Information Processing Society Conference Korea Information Processing Society*, 581-584.

[32] Ve, S., Shin, C., & Cho, Y. (2021). Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city. *Building Research & Information*, *49*(1), 127-143. https://doi.org/10.1080/09613218.2020.1809983

[33] Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*, 353-366. https://doi.org/10.1016/j.comcom.2020.02.007

[34] Ve, S., Park, J., & Cho, Y. (2020). Seoul bike trip duration prediction using data mining techniques. *IET Intelligent Transport Systems*, *14*(11), 1465-1474. https://doi.org/10.1049/iet-its.2019.0796

[35] Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, *140*, 81-97. https://doi.org/10.1016/j.enbuild.2017.01.083

[36] Ve, S. & Cho, Y. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, *53*(1), 166-183. https://doi.org/10.1080/22797254.2020.1725789

[37] Sathishkumar, V. E. & Yongyun, C. (2021). *A Study on Regression Accuracy improvement using hybrid feature selection method*. Sunchon National University Graduate School.

Contact information:

Sathishkumar V. E., Postdoctoral Researcher
Hanyang University, Department of Industrial Engineering,
222 Wangsimini-ro, Seondong-gu, Seoul, Republic of Korea, 04763
E-mail: sathishkumar@hanyang.ac.kr

Yongyun CHO, Professor
(Corresponding author)
Sunchon National University, Department of Information and Communication Engineering, Suncheon, Republic of Korea
E-mail: yycho@scnu.ac.kr