International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Social Spider Optimization Algorithm for Effective Data Classification: An Application of Stock Price Prediction

R. Saravanan, P. Sujatha, G. Kadiravan, J. Uthayakumar

Abstract: Presently, data classification has become a hot research topic, intends to categorize the data into a predefined number of classes. The data classification problem has been considered as a NP hard problem and various optimization algorithms are introduced to solve it. Social spider optimization (SSO) algorithm is originally developed to resolve continuous and optimization problems. In line with, it has been altered to manage different kinds of optimization process and also to be employed for data investigation. And, some other studies are also investigated the use of SSO algorithm in different domains. In this paper, we introduce new SSO algorithm particularly applicable for data classification process. In order to validate the performance of the SSO algorithm, a real time problem of stock price prediction (SPP) is employed. For experimentation, the results are validated by testing the SSO algorithm against four datasets such as Dow Jones Index (DJI) dataset, three own datasets gathered from Yahoo finance on the basis of daily, weekly and yearly. The empirical result states that the proposed algorithms perform well and it is noted that the classification performance of the SSO algorithm better than the compared methods.

Keywords :Social Spider algorithm; Optimization algorithm; Classification; Stock price prediction

I. INTRODUCTION

Data classification is a viral part in the data mining process, which is employed to the databases for labeling the new input objects whose labels are not known [1]. The data classification techniques aim to extract the relationship between a collection of feature variables and a target variable of interest. As various real-world problems are often defined by the relationship between the available features and target variable, it achieves a wide range of applicability. In general, the data classification process needs a classifier model, attained from a sample of objects from the database where the labels are already present. The procedure of construction such kind of model is called learning or training and it refers to the parameter changes of the algorithm for maximizing the generalization ability [2].

The prediction for unlabeled data is carried out by the models created in the training process or few of the lazy learning processes that has the capability of comparing objects classified previously with the new objects needs to be classified. The outcome of a classification method may be applied for a test instance in any of the two ways namely discrete label and numerical score. In the former type, a label will be assigned for the test instances. In the latter type, a numerical score will be provided to every class label and test instance combination. It is noted that the numeric score can be transformed to a discrete label for a test instance, by choosing the class with the high score for that test instance. The benefit of a numerical score is that providing a possibility for comparing the relativity of various test instances which belongs to a specific class of importance, and rank them under requirement. Data classification is a supervised learning approach where the machine learns something from the provided input data and then employs the learning process for classifying the new observation. This dataset might contain only two class labels such as identifying the mail as spam or not, and sometimes it might contain many classes. Various application areas of data classification problems are speech recognition, handwriting recognition, bio metric identification, document classification etc. [3]

The stock price prediction is also an important data classification problem which aims to classify the stocks as profit or loss, i.e. stock prices rises or not. Stock is a crucial investment in financial sectors and presently it seeks much attention among the research community. Investors and research people are considering multidimensional factors to suggest a methodology that will choose the stock precisely with the guaranteed profit in return. Due to the diverging nature of SPP, it is diligent to predict stock prices each and every business day. An accurate SPP can lead the investors to gain higher profit while exchanging stocks. Hence unambiguously, SPP has become more popular and highly helpful for investors [4]. A study [5] revealed that a proper SPP can lead the trading to a more profitable way and it also recommended that researchers should concentrate on precisely identifying the direction of movement and to reduce the variations from the real observed values.

Manuscript published on 30 September 2019
* Correspondence Author
R. Saravanan*, Department of Computer Science, Pondicherry University, Puducherry, India, r.saravanan26@gmail.com
P. Sujatha, Department of Computer Science, Pondicherry University, Puducherry, India, spothula@gmail.com
G. Kadiravan Department of Computer Science, Pondicherry University, Puducherry, India, kathirkadir@gmail.com
J. Uthayakumar, Department of Computer Science, Pondicherry University, Puducherry, India,uthayresearchscholar@gmail.com

The precise SPP will be a huge benefit to the investors [6]. But, the performance of stock exchange is based on several qualitative factors like politics, economics, weather, etc. The stock markets rapidly changing and they show more variation. This nonlinear and complex nature of stock market makes SPP more difficult [7, 8]. The prediction of future stock prices is defined by factors like data intensity, noise, non-stationary, random nature, uncertainties and invisible interlinks [9–11].

Generally, there are two classifications of SPP which includes the SPP movement and predicting the value of the stock price. The first type is considered as a classification problem and the other is treated as a regression problem. But both depend on auto regression and multi-variable regression models. The auto regression model treats the problem based on time sequences and the sequences are partitioned to various segments. The partitioned segments are employed as raw data for SPP. In multi-variable regression model, the technical variables are chosen as raw data for SPP.

In recent years, many methods have been proposed to predict the trends in stock price variations. Prediction of upcoming stock price using the existing financial information is highly valuable for investors. Investors like to predict the increasing or decreasing trends in stock price over a period of time. Prior to making an investment decision in a firm, a prediction method was used to foresee stock prices using the existing and present financial data about the firm. Financial balance sheets and several ratios represent the status of company and it is the fundamental for technical investigation.

The experts applied various mathematical models using the previous data to validate company's intrinsic value like Graham number. Graham number and Graham's criterion is assumed as a popular method of SPP [12]. Because of increasing instability in the present market scenario, it is very hard to identify a company which satisfies Graham's principles on current situation. Due to these modifications, the requirement for adjusted models started. In addition, the stock market trends change over time [13, 14]. Generally, Machine Learning (ML) technique handles SA as a traditional classification problem. Hybrid methods integrate the features of lexicon based and ML methods (Naive Bayes (NB), Bayesian Network (BN), Maximum Entropy (ME), Support Vector Machine (SVM), Neural Network (NN), Decision Tree (DT), etc.) are also applied to solve SPP classification problem. Metaheuristic approaches are also used for SPP due to the nature of effective searching strategies. Some of the meta-heuristic algorithms are ant colony optimization (ACO), particle swarm optimization (PSO), genetic algorithm (GA), artificial bee colony (ABC) optimization, bat algorithm (BA) and so on. Social spider is one of the meta-heuristic algorithm [15], derived from the foraging nature of social spider. SSO algorithm is used in several application areas like electrical engineering, clustering, images processing applications, etc.

SSO algorithms are employed to different traditional optimization problems. In [16], a SSO algorithm is presented for global optimization problems. The inflation loss mechanism is used to avoid the premature convergence of SSO algorithm. A SSO algorithm for handling constrained optimization problems is presented in [17]. A hybridization of SSO algorithm is derived using simplex Nelder-Mead technique [18] to solve the integer programming and min-max problems. The SSS algorithm combines the positive capability to do the exploration and exploitation process whereas the Nelder-Mead method improves the best obtained solution from SSO algorithm. It is intended to enhance the searching capability and avoids the execution of the algorithm with more number of iterations with no improvement. A binary social spider (BSS) algorithm is introduced for knapsack KP01 problem [19] effectively. In [20], the SSO algorithm is applied to resolve the non-convex problems of Economic Load Dispatch (ELD). A modification in SSO algorithm is made and developed a MSSO algorithm [21] for the improving the exploration process of the continuous optimization problem. In [22], the SSO algorithm is integrated to GA to develop a HSSOGA for simplifying the energy function of the molecule. For clustering the text documents, the SSO algorithm is introduced in [23]. In this paper [24], a simplex method-based social spider optimization (SMSSO) algorithm is presented to attain effective clustering process. The SSO algorithm is employed in image processing applications. An image fusion approach using SSO algorithm for image enhancement is presented in [25]. In [26], a histogram based bi-modal and multi-modal thresholding is developed for grayscale images by the use of SSO algorithm.

In this paper, we introduce new SSO algorithm for data classification problem to identify the classes for every applied instance. The SSO algorithm is derived for the data classification process to classify the data in an effective way. In order to validate the performance of the SSO algorithm, a real time problem of stock price prediction is employed. For experimentation, the results are validated by testing the SSO algorithm against four datasets such as Dow Jones index (DJI) dataset, three own datasets gathered from Yahoo finance on the basis of daily, weekly and yearly. The empirical result states that the proposed algorithms perform well and it is noted that the classification performance of the SSO algorithm is better than the other methods.

The paper organization is given as follows: section 2 gives an overview of SSO algorithm. The proposed SSO algorithm is explained in Section 4. The experimental analysis and conclusions are provided in Section 4 and Section 5 respectively.

II. BIOLOGICAL NATURE OF SOCIAL SPIDERS

Social insect society is a complicated distributed system which self organizes themselves in a collection of limitations. They are found to be efficient in the manipulation and exploitation of the environment, defends the food resources and new births, and also allows task specialization between the members of the group. A social insect colony operates as a combined model which does not possess the capability to function in a distributed way. It is significant to recognize that global order in social insects arises due to the implicit interaction between the members. Some species of spiders exhibit a different degree of social behaviors of spiders [27]. The nature of spiders is divided into two types namely solitary spiders and social spiders [28].

This classification is made depending upon the cooperative nature which they demonstrate [29]. The solitary spider creates and maintains their own web when lives in a scarce contact to other individuals of the identical species. Contrastingly, the social spiders generate colonies which remain together in a communal web with closer relationship to the group members [30]. The colony of the social spider includes two basic elements namely members and communal web. The members of the colony consist of male and female spiders and it is a fascinating feature that the female spider population is much higher than the male population. In [28], it is reported that the male population rarely reaches to 30 percent of the entire colony population. Every member in the colony exhibits some actions based on the gender like constructing and preserving the communal web, attacking preys, mating and communication. The communication takes place between the spiders takes place in an intrinsic as well as extrinsic way [32]. The direct explicit communication indicates body contact or the exchanging fluid like mating. The intrinsic communication uses the communal web as a communication medium which transfers the information to every colony member exists in the web. The information will be transmitted using smaller vibrations which is an important dimension for the cooperative decision between the members. These vibrations are used by the colony members for decoding different information like size of the prey trapped in the web, features of the nearby members, etc.

The intensity of the vibrations is based on the weight and distance of the spiders which has generated them. In spite of the complexity, every cooperative global pattern in the colony level is produced from the interaction between the colony members [33]. This intrinsic interaction includes a collection of easier behavioral rules followed by every spider in the colony. The behavioral rules are partitioned into social interaction (cooperative behavior) and mating. Based on the gender of the spiders, the spiders interact with each other. The female spiders depict an affinity to interaction through attraction or dislike over others, irrespective of the gender.

A specific female spider shows attraction or repulsion over others based on the emitted vibrations in the communal web and indicates stronger colony members. As the vibrations are mainly based on the weight and distance of the members that creates it, intensity of the vibrations will be stronger in case of big or nearby spiders. When the spider is very big, there are more chances it will be treated as a colony member. The last choice of attraction or repulsion to a fixed member based on the internal state has strong influence over different factors like reproduction cycle, curiosity, and other random phenomenon. In contrast to female one, the nature of males is based on reproduction. They identify themselves as a sub part of alpha males which overlook the colony resources. Consequently, the male spiders are portioned to dominant and non-dominant ones. The dominant male spiders have best fitness features compared to non-dominant spiders. Typically, the dominant male spiders show attraction towards the nearest female ones in the communal web. Contrastingly, the non-dominant male spiders show a tendency to focus on the center point of the male spiders as a procedure to take over the resources wasted by the dominant ones.

And, mating is an important process which is not only used for colony survival, and also for communication between the colony members. It is usually done between the dominant male and female spiders [34]. In those situations, if a Fig. 1.

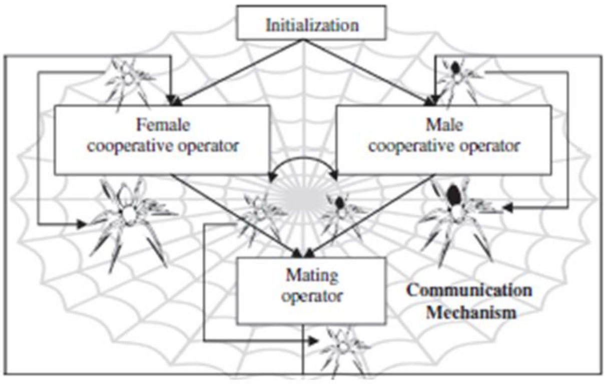

Fig.1 Workflow for SS Algorithm

Workflow of SS algorithm dominant male finds one or many female spiders in a particular area, it undergoes mating with every female present in the area and produces offspring [35].

III. PROPOSED DATA CLASSIFICATION ALGORITHM

The proposed SSO based data classification operates as follows: the female spiders in the colony members determine the class label based on the attributes hold by the male spiders. The male spiders indicate the value of the attribute using the intensity of the vibration. Based on the intensity of the vibration reached by the female spider, the attribute value will be determined and decides the class label for every instance. The pseudo code of the SSO algorithm for data classification is given in Algorithm I and the variables used are given in Table 1. The SSO algorithm considers the whole search space as a communal web, where each social spider interacts with one another. Every solution in the search space indicates the position of the spider in the communal web. Each spider will receive a weight based on the fitness value of the solution that is defined by the social-spider. The SSO algorithm defines two types of search agents namely male spiders and female spiders. The overall operation of the SSO algorithm is illustrated in Fig.1, which comprises of four blocks namely population initialization, female cooperative operator, male cooperative operator and mating operator.

A. Initialization

The SSO algorithm starts with the initialization of the population of N spider positions (solutions). The whole population contains female f_i and male m_i spiders. The number of f_i (N_f) is arbitrarily selected between 65-90% and is determined by Eq. (1) and the number of $m_i(N_m)$ is also determined by Eq. (2):

$$N_f = \mathrm{floor}[(0.9 - \mathrm{rand}(0,1).\, 0.25).\, N] \quad (1)$$

$$N_m = N - N_f \quad (2)$$

The position of the f_i is randomly generated between the lower and upper initial parameter bound p^{low} and p^{high} which are represented by

$$f_{i,j}^{0} = p_j^{low} + rand(0,1).(p_j^{high} - p_j^{low}) \quad [3]$$

where $i = 1, 2, \dots, N_f; j = 1, 2, \dots, n$.

Then, the position of the m_i is also randomly generated and equated as

$$m_{i,j}^{0} = p_j^{low} + rand(0,1).(p_j^{high} - p_j^{low}) \quad (4)$$

where $i = 1, 2, \dots, N_m;\ j = 1, 2, \dots, n$.

B. Fitness assignation

Biologically, the size of the spiders is the main feature that is used to evaluate spider's capacity to accomplish the allocated process. In the proposed model, each spider receives a weight w_i that indicates the solution quality of the spider i (irrespective of gender) of the population S. The weight of the each spider is determined as

$$\frac{J(s_i) - worst_s}{best_s - worst_s} \quad (5)$$

where $J(s_i)$ is the fitness value, $worst$ and $best$ is the maximum and minimum values of the solution. The values $worst_s$ and $best_s$ are represented using the Eq. (6)

$$best_s = \max_{k\in\{1,2,\dots,N\}}(J(s_k)) \text{ and}$$

$$worst_s = \min_{k\in\{1,2,\dots,N\}}(J(s_k)) \quad (6)$$

C. Vibration Modelling

The communal web is treated as a communication medium to exchange information between the members of the colony. This information is coded as small vibrations which are important for the cooperative operation of every individual in the population. The vibrations are based on the weight and distance of the spider which has produced it. As the distance is related to the individuals that generates the vibrations and the colony members are present in the nearby distance will receive stronger vibrations and vice versa. To information exchange between the colony members i and j, the vibration can be mathematically defined as follows:

$$Vib_{i,j} = w_j e^{-d_{i,j}^2} \quad (7)$$

where $d_{i,j}$ is the Euclidian distance between two colony members i and j. Using these vibrations, the attribute value of the instance will be sent by the member i to member j. There are three kinds of vibrations takes place between i and j and they are represented as $Vibc_i$, $Vibb_i$ and $Vibf_i$.

Vibrations $Vibf_i$ is received by the individual i (si) as a result of the information sent by the member c (sc) which is closer to i as well as with higher weight compared to i (wc > wi).

$$Vibc_i = w_c e^{-d_{i,c}^2} \quad (8)$$

The vibrations $Vibb_i$ received by the individual i as a result of the information transmitted by the member b (sb) which has best weight (best fitness value) of the whole entire population S,

$$Vibb_i = w_b e^{-d_{i,b}^2} \quad (9)$$

Finally, $Vibf_i$ defines the information transmitted from the member i to the nearest female individual $f(s_f)$ can be represented as

$$Vibf_i = w_f e^{-d_{i,f}^2} \quad (10)$$

D. Female cooperative operator

The f_i shows an attraction or repulsion to other spiders irrespective of gender. The movement of attraction or repulsion depends on is based on different random criteria. A uniform random number r_m is generated in the range of [0, 1]. When r_m is lesser than a predefined threshold PF, an attraction is generated else a repulsion is created and is given in Eq. (11).

$$f_i^{t+1} = \begin{cases} f_i^t + \alpha.Vibc_i.(s_c - f_i^t) + \beta.Vibb_i.(s_b - f_i^t) + \delta.(rand - 0.5) & \text{with probability PF} \\ f_i^t - \alpha.Vibc_i.(s_c - f_i^t) - \beta.Vibb_i.(s_b - f_i^t) + \delta.(rand - 0.5) & \text{with probability } 1 - PF \end{cases} \quad (11)$$

E. Male cooperative operator

The m_i has a weight value more than the median of the N_m is called as dominant D and the remaining m_i are called as non-dominant ND. The median weight is indexed by N_{f+m}. The location of the m_i can be equated as

$$m_i^{t+1} = \begin{cases} m_i^t + \alpha.Vibf_i.(s_f - m_i^t) + \delta.(rand - 0.5) & \text{if}(w_{N_{f+i}} > w_{N_{f+m}}) \\ m_i^t + \alpha.\left(\frac{\sum_{h=1}^{N_m} -h-N_f+n}{\sum_{h=1}^{N_m} .w_{N_f}+h} - m_i^t\right) \end{cases} \quad (12)$$

F. Mating operator

In general, mating is performed between D and f_i in case of a f_i is found by D in a specific range, the mating range is equated as

$$r = \frac{\sum_{j=1}^{n}(p_j^{high} - p_j^{low})}{2.n} \quad (13)$$

The spider with more weight has higher probability of generating the offspring.

Nomenclature
$D() \leftarrow$ Dataset
$C_V \leftarrow$ Continuous Variables
$D_V \leftarrow$ Discrete Variables
$CS_V \leftarrow$ Class
$f() \leftarrow$ Objective Function
$TN \leftarrow$ True Negative
$TP \leftarrow$ True Positive
$FP \leftarrow$ False Positive
$FN \leftarrow$ False Negative

Algorithm I: Pseudo code of SSO algorithm

Input: Dataset ($D\left(C_V, D_V, \overrightarrow{CS_V}\right)$), $f()$
Define C_V, D_V, CS_V from D
// Discretization Factor for Continuous Variables
//Fixed Frequency Distribution (FFD) [1]
$DC_V \leftarrow C_V$ using FFD
Initialize the Parameters $c, r_a, N \leftarrow$ Number of Spiders,
$d \leftarrow$ Dimension of each spider, $t \leftarrow 1$
$N \leftarrow N_M + N_F$
// Initialize the Operators for Discrete & Discretized
Continuous Variables
$OD_V \leftarrow [=, \neq, \emptyset]$
$ODC_V \leftarrow [\leq, \geq, \emptyset]$
//Population Initialization
For $\forall i \in N$ **do**
For $\forall i \in d$ **do**
$PopV_{i,j} \leftarrow randi(1, |DC_V|, 1)$
$PopO_{i,j} \leftarrow randi(1, |ODC_V|, 1)$
End For
$Pop_i \leftarrow [PopV_i \; PopO_i]$
End For

//Fitness Calculation
For $\forall i \in N$ **do**
$\overrightarrow{PC} \leftarrow Classification\left(Pop_i, \overrightarrow{CS_V}\right)$
$Fit_i \leftarrow f\left(\overrightarrow{PC}\right)$
End For
Repeat
//Calculation of Mating Radius
For $\forall i \in N$ **do**
$$MR_i \leftarrow \frac{\sum_{e=1}^{N_m} m_e(o).W_{N_{f+h}}}{\sum_{e=1}^{N_m}.W_{N_{f+h}}}$$
End For
//Calculation of weights of Each Spider
For $\forall i \in N$ **do**
$$W_i = \frac{fit(P) - worst_p}{best_p - worst_p}$$
End For
//Calculation of Vibration of Female Spiders
For $\forall i \in N_F$ **do**
$$FV_i = \begin{cases} FV_i(t) + x.V_{i,c}.(P_c - FV_i(t)) + y.V_{i,d}.(P_d - FV_i(t)) \\ \qquad + z.\left(rand - \frac{1}{2}\right) \eta < \gamma \\ FV_i(t) - x.V_{i,c}.(P_c - FV_i(t)) - y.V_{i,d}.(P_d - FV_i(t)) \\ \qquad + z.\left(rand - \frac{1}{2}\right) \eta \geq \gamma \end{cases}$$
End For
For $\forall i \in N_M$ **do**
$MV_i =$
$$\begin{cases} MV_i(t) + x.V_{i,f}.\left(P_f - MV_i(t)\right) + +z.\left(rand - \frac{1}{2}\right) \\ \qquad if\, W_{N_{f+i}} > W_{N_{f+m}} \\ MV_i(t) + x.\left(\frac{\sum_{e=1}^{N_m} m_e(t).W_{N_{f+h}}}{\sum_{e=1}^{N_m} W_{N_{f+h}}} - m_i(t)\right) \\ \qquad if\, W_{N_{f+i}} \leq W_{N_{f+m}} \end{cases}$$
End For
//Mating Process and Fitness Calculation
$t \leftarrow t + 1$
Until (termination condition satisfied)
OUTPUT:$min\left(f(Pop)\right)$, Optimized Rule Set

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To ensure the goodness of the proposed SSO algorithm, it is applied to the problem of SPP. The dataset, measures and the attained results are discussed in the upcoming subsections.

A. Dataset description

For precise SPP, the dataset is collected by the past history of stock prices of four dataset with massive instances. The datasets of SPP are namely Dow Jones Index(DJI) [36], YahooFinance_2007_2017_Daily,YahooFinance_2007_2017_Weekly,YahooFinance_2007_2017_Monthly[37] are taken. The particulars of the dataset are demonstrated in Table 2. The frequency distribution of these dataset is shown in Figs. 2-5. For SPP dataset, the entire count of instances are 4210 and count of the distinct dataset such as DJI YahooFinance_2007_2017_Daily,YahooFinance_2007_2017_Weekly, Yahoo Finance_ 2007_2017_Monthly are 750, 2755, 572, 133 correspondingly. For DJI dataset, among the total number of instances, there are 375 instances for both increased and decreased in stock prices. For Yahoo Finance_2007_2017_Daily, there are 1269 instances for increase stock prices and 1486 for decreased stock prices. For Yahoo Finance_2007_2017_Weekly, there are 247 increased stock prices and 325 decreased stock prices. For Yahoo Finance_2007_2017_Monthly, there are 48 instances which are increased stock prices and 85 which are decreased stock prices. Among these datasets, Yahoo Finance_2007_2017_Daily has large number of increased and decreased stock rates.

Table 2 Dataset Description

S. No	Dataset	No. of Instances	No. of Features	No. of Classes	↑↓
1	Dow Jones Index	750	16	2	375/375
2	YahooFinance_2007_2017_Daily	2755	7	2	1269/1486
3	YahooFinance_2007_2017_Weekly	572	7	2	247/325
4	YahooFinance_2007_2017_Monthly	133	7	2	48/85

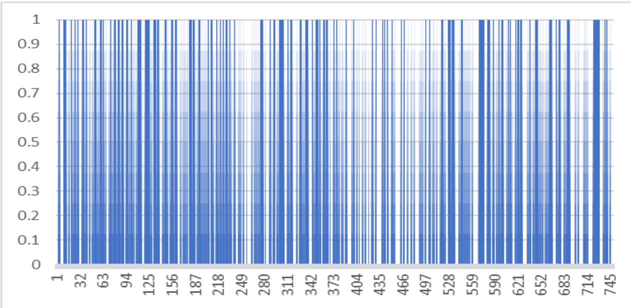

Fig. 2. Frequency distribution of Dow Jones Index (DJI) Dataset

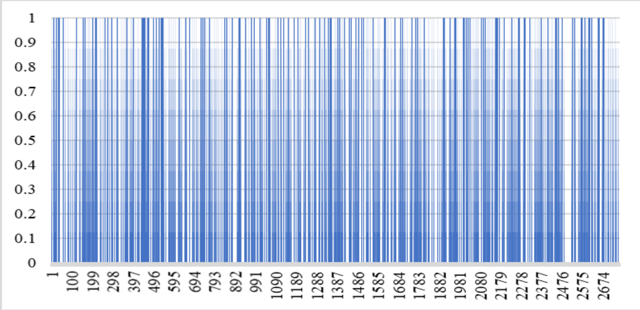

Fig. 3. Frequency distribution of Yahoo Finance _2007 _2017 _Daily Dataset

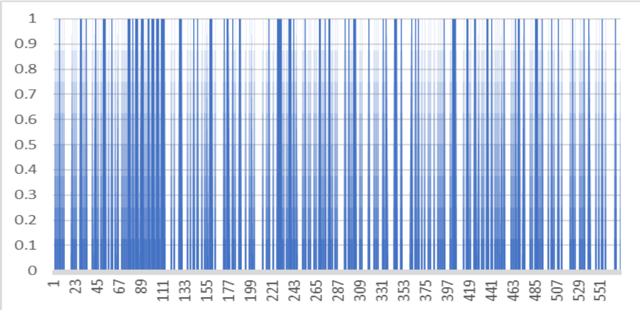

Fig. 4. Frequency distribution of Yahoo Finance _2007 _2017 _Weekly Dataset

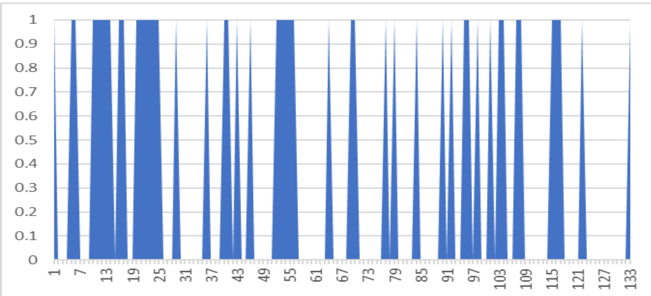

Fig. 5. Frequency distribution of Yahoo Finance _2007 _2017 _Monthly Dataset

B. Performance Measure

To assess the performance of SSO algorithm against four dataset, 15 measures is used which are generated from a 2x2 confusion matrix. It is generally a 2x2 matrix and contains 4 elements related to real and predicted classes. The elements in confusion matrix are True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) values. The overall performance of the model is determined using the data present in the matrix.

C. Results Analysis

Dow Jones Index (DJI) dataset:

Table 3 and Table 4 provide the confusion matrix and the classification results attained by the SSO algorithm. Among the total 750 instances in DJI dataset, 357 instances are correctly classified to increased stock prices and 343 instances are correctly classified into decreased stock prices by the SSO algorithm. With respect to the classification performance, the results of the proposed algorithms are ensured by comparing its results with six other classifiers namely ACO, XGBoost, OlexGA, MLP, RBF and RF. Regarding the classifier results, DL-TLBO model is verified by comparing its results with three classifiers include DL, LR and RBF models.

Table 4 values show that the proposed SSO algorithm attains a minimum FPR and FNR values of 7.29 and 6.05values respectively. At the same time, it is noticed that the Olex-GA shows worst performance with a highest FPR value of 50.00. In addition, it is observed that the ACO, XGBoost and RG classifiers manages to perform well than Olex-GA, MLP and RBF classifiers. But they showed higher FPR and FNR values over the SSO algorithm.

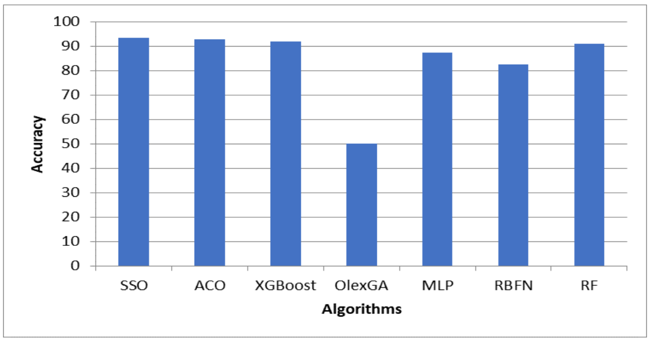

Fig. 6. Comparison of Various Methods in terms of Accuracy (DJI Dataset)

Table 3 Confusion matrix for Dow Jones Index (DJI) Dataset

Experts	SSO		ACO		XGBoost		OlexGA		MLP		RBF		RF	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
↑	357	27	350	25	348	27	0	375	318	57	310	65	345	30
↓	23	343	30	345	35	340	0	375	39	336	67	308	37	338

Table 4 Comparison of proposed method with existing methods for Dow Jones Index (DJI) Dataset

Classifier	FPR	FNR	Sens.	Spec.	Accu.	F-score	Y	ρ^+	ρ^-	DP	G-measure	MCC	FDR	FOR	Kappa
SSO	07.29	06.05	93.94	92.70	93.33	93.45	86.65	15.31	12.87	3.01	93.45	0.86	07.03	06.28	86.66
ACO	06.75	07.89	92.10	93.24	92.66	92.71	85.34	11.81	13.63	2.71	92.71	0.85	06.66	08.00	85.33
XGBoost	07.35	09.13	90.86	92.64	91.73	91.82	83.50	10.13	12.35	2.54	91.82	0.83	07.20	09.33	83.46
OlexGA	50.00	-	-	50.00	50.00	0	-	-	-	-	-	-	100	0	0
MLP	14.50	10.92	89.08	85.49	87.20	86.89	74.57	07.82	06.14	2.29	86.91	0.74	15.20	10.40	74.40
RBFN	17.42	17.77	82.22	82.57	82.40	82.44	64.80	04.64	04.71	1.69	82.44	0.64	17.33	17.86	64.80
RF	08.15	09.68	90.31	91.84	91.06	91.14	82.16	09.48	11.07	2.47	91.15	0.82	08.00	09.86	82.13

Yahoo Finance _2007 _2017 _Daily Dataset

The confusion matrix and the classification results obtained from them are given in Table 6 and Table 5. From the total of 2755 instances in Yahoo Finance daily dataset, 396 instances are correctly classified to increased stock prices and 1360 instances are correctly classified into decreased stock price. In contrast, the number of instances correctly classified to increased and decreased stock is 565 and 610 respectively.

Based on the classification performance, the obtained values are the comparative results are tabulated in Table 5. As said earlier, the value of FPR, FNR, FOR and FDR should be as low as possible. From the table, it is showed that the FPR, FNR, FOR and FDR of SSO algorithm are 39.09, 24.13, 8.47 and 68.73 respectively. It is observed that the poor performance interms of FPR, FNR, FOR and FDR are attained by XGBoost, Olex-GA, RBF and MLP. In continuity, the value of the classification accuracy is significantly higher for SSO algorithm which attains an accuracy of 63.73. From the Fig. 7, it is clear that the order of increased classification performance is SSO, ACO, XGBoost, Olex-GA, MLP, RBF and RF respectively. In addition, in terms of sensitivity and specificity, the SSO algorithm achieves maximum values of 75.86 and 60.9 respectively.

Contrastingly, the RBF attains a lowest sensitivity and specificity values of 43.1 and 53.87 respectively. In line with, the XGBoost, RBF and RF classifiers showed competitive performance and it fails to outperform SSO algorithm. Similarly, the SSO algorithm reported an F-score of 44.22 which is higher than the values obtained by the compared methods. Next, the F-score attained by XGBoost and RF classifiers are better with the values of 47.71 and 46.23 respectively. The kappa value of SSO algorithm is 23.74 indicating effective classification results. At the same time, RBF classifier attains the negative performance with the F-score of -0.269. Though the RF and XG-Boost classifiers outperform the compared methods, it also attains inferior results to the SSO algorithm.

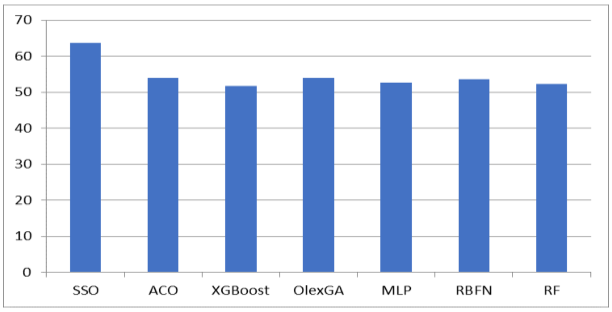

Fig. 7. Comparison of Various Methods in terms of Accuracy (Yahoo Finance Daily Dataset)

Table 5 Confusion matrix for the Yahoo Finance _2007 _2017 _Daily dataset

Experts	SSO		ACO		XGBoost		OlexGA		MLP		RBF		RF	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
↑	357	27	350	25	348	27	0	375	318	57	310	65	345	30
↓	23	343	30	345	35	340	0	375	39	336	67	308	37	338

Table 6 Comparison of proposed method with existing methods for Dow Jones Index (DJI) Dataset

Classifier	FPR	FNR	Sens.	Spec.	Accu.	F-score	Y	ρ^+	ρ^-	DP	G-measure	MCC	FDR	FOR	Kappa
SSO	39.09	24.13	75.86	60.90	63.73	44.22	36.76	2.52	1.94	1.141	48.65	0.28	68.79	8.47	23.74
ACO	45.51	49.50	50.49	54.48	54.04	19.36	4.98	1.10	1.10	0.06	24.59	0.031	88.02	10.02	2.06
XGBoost	44.67	52.32	47.67	55.32	51.79	47.71	3.00	1.05	1.06	-0.02	47.71	0.03	52.24	44.75	3.00
OlexGA	46.06	-	-	53.93	53.94	0	-	-	-	-	-	-	100	0	0
MLP	45.99	53.63	46.36	54.00	52.66	25.48	0.36	1.0068	1.0079	-0.08	28.54	0.0028	82.42	17.36	0.221
RBFN	46.12	36.89	43.10	53.87	53.64	3.76	-3.02	0.94	0.93	-0.18	9.21	-0.0087	98.02	2.22	-0.269
RF	44.55	51.91	48.08	55.44	52.30	46.23	3.52	1.06	1.07	-0.006	46.26	0.035	55.47	41.04	3.49

Yahoo Finance _2007 _2017 _Weekly Dataset

The confusion matrix and the classification results obtained from them are given in Table 7 and Table 8. From the total of 572 instances in Yahoo Finance weekly dataset, 223 instances are correctly classified as increased stock prices.Table 8 shows the attained classification results of various methods under several performance metrics for yahoo finance weekly dataset in the year 2007 to 2017. For Weekly dataset, the proposed SSO algorithm produces best results and attained a minimum FPR and FNR value of 18.63 and 25.2 respectively. Among the compared methods, RBF classifier showed poor performance with a highest FPR and FNR of 44.92 and 59.47 respectively. At the same time, ACO manages to perform well and showed lower FPR and FNR value than compared methods, it showed inefficiency over SSO algorithm. Interms of accuracy, Fig. 8 depicts the comparative results of different classifiers on the Yahoo Finance Weekly dataset. From the Fig., it is shown that order of maximum classification accuracy attained by the different classifiers is SSO, ACO, Olex-GA, XG-Boost, RBF, MLP and RF respectively. In the same way, the SSO classifier showed maximum classification performance with an accuracy of 78.49 which is higher than all the compared models. On comparing the classifiers based on kappa value, the results reported that the RBF and RF classifiers showed poor classification results with the negative kappa value of -0.108 and -4.333 respectively. However, the SSO algorithm tries to show better performance with a kappa value of 56.24. From the Table 8 and Fig. 8, it is reported that the classification results based on Yahoo Finance weekly dataset verified that the SSO algorithm is found to be efficient than the compared classifiers.

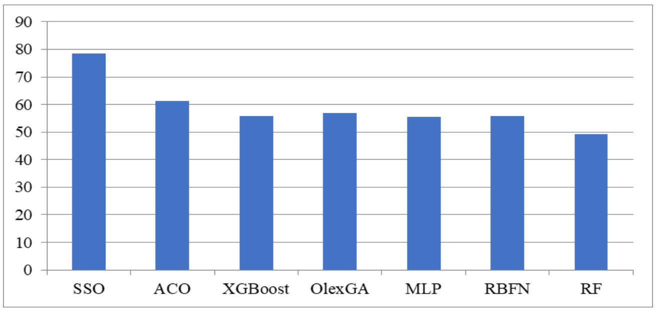

Fig. 8. Comparison of Various Methods in terms of Accuracy (Yahoo Finance Weekly Dataset)

Table 7 Confusion matrix for the Yahoo Finance _2007 _2017 _Weekly Dataset

Experts	SSO		ACO		XGBoost		OlexGA		MLP		RBF		RF	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
↑	187	60	116	131	103	144	0	247	24	223	18	229	92	155
↓	63	262	91	234	109	216	0	325	31	294	24	301	135	190

Retrieval Number: C4603098319/19©BEIESP
DOI:10.35940/ijrte.C4603.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

Table 8 Comparison of proposed method with existing methods for YahooFinance_2007_2017_Weekly Dataset

Classifier	FPR	FNR	Sens.	Spec.	Accu.	F-score	Y	ρ^+	ρ^-	DP	G-measure	MCC	FDR	FOR	Kappa
SSO	39.09	24.13	75.86	60.90	63.73	44.22	36.76	2.52	1.94	1.141	48.65	0.28	68.79	8.47	23.74
ACO	45.51	49.50	50.49	54.48	54.04	19.36	4.98	1.10	1.10	0.06	24.59	0.031	88.02	10.02	2.06
XGBoost	44.67	52.32	47.67	55.32	51.79	47.71	3.00	1.05	1.06	-0.02	47.71	0.03	52.24	44.75	3.00
OlexGA	46.06	-	-	53.93	53.94	0	-	-	-	-	-	-	100	0	0
MLP	45.99	53.63	46.36	54.00	52.66	25.48	0.36	1.0068	1.0079	-0.08	28.54	0.0028	82.42	17.36	0.221
RBFN	46.12	36.89	43.10	53.87	53.64	3.76	-3.02	0.94	0.93	-0.18	9.21	-0.0087	98.02	2.22	-0.269
RF	44.55	51.91	48.08	55.44	52.30	46.23	3.52	1.06	1.07	-0.006	46.26	0.035	55.47	41.04	3.49

Yahoo Finance _2007 _2017_Monthly Dataset:

The classification results and confusion matrix obtained from them are given in Table 9 and Table 10. From the total of 133 instances in Yahoo Finance Monthly dataset, 34 instances are correctly classified to increased stock prices whereas the ACO, MLP and RF classifiers properly identifies 30, 21 and 17 instances respectively. Table 9 shows the attained classification results of various methods under several performance metrics for yahoo finance monthly dataset in the year 2007 to 2017. Similar to weekly dataset, the table shows that the highest level of accuracy 87.96 is achieved by SSO which is higher than the ACO algorithms Though Olex-GA attains an accuracy of 91.72, it fails to achieve better results than SSO and ACO algorithms. Similarly, the F-score value of the SSO algorithm are found to be high compared to the compared classifiers. The ACO algorithm obtains an F-score value of 73.17 which is higher than the compared methods except SSO algorithm. Based on sensitivity and specificity, the SSO algorithm obtains maximum values of 94.44and 85.56respectively. And, the RBF attains a lowest sensitivity and specificity values of 27.27 and 63.11 respectively. In line with, the XGBoost, MLP and RF classifiers showed competitive performance and it fails to outperform ACO, SSO algorithm. Since the kappa value closer to 100 implies better classification performance where the SSO algorithm attains a maximum kappa value of 72.42. At the same time, it can also be seen that the XGBoost and RBF classifiers achieved worst classification with the negative kappa values of -2.108 and -3.799 respectively. From the above tables and graphs, it is evident the SSO algorithm outperforms the compared methods in a significant way. Specifically, the use of SSO classifier shows the superiority with easier training process. Finally, it is concluded that the SSO algorithm is an effective classifier and is verified on SPP model over the state of art methods.

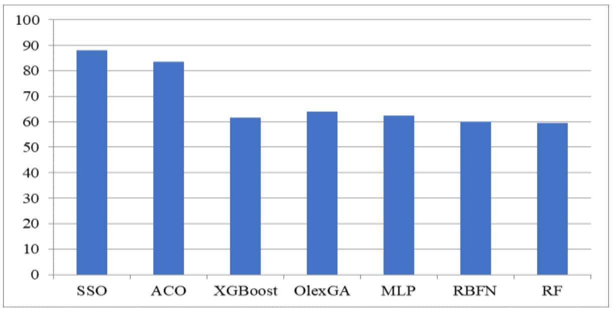

Fig. 9. Comparison of Various Methods in terms of Accuracy (Yahoo Finance Monthly Dataset)

Table 9. Confusion matrix for the Yahoo Finance _2007 _2017 _Monthly Dataset

Experts	SSO		ACO		XGBoost		OlexGA		MLP		RBF		RF	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
↑	34	14	30	18	2	46	0	48	21	27	3	45	17	31
↓	2	83	4	81	5	80	0	85	23	62	8	77	23	62

Table 10. Comparison of proposed method with existing methods for YahooFinance_2007_2017_Weekly Dataset

Classifier	FPR	FNR	Sens.	Spec.	Accu.	F-score	Y	ρ^+	ρ^-	DP	G-measure	MCC	FDR	FOR	Kappa
SSO	14.43	5.55	94.44	85.56	87.96	80.95	80.01	15.40	6.54	3.06	81.79	0.74	29.16	2.35	72.42
ACO	18.18	11.7	88.23	81.81	83.45	73.17	70.05	6.95	4.85	2.17	74.26	0.63	37.5	4.70	61.71
XGBoost	36.50	71.42	28.57	63.49	61.65	7.27	-7.93	0.889	0.7826	-0.5699	10.91	-0.0369	95.83	5.88	-2.108
OlexGA	36.09	-	-	63.90	63.90	0	-	-	-	-	-	-	100	0	0
MLP	30.33	52.27	47.72	63.66	62.40	45.65	17.39	1.33	1.57	0.10	45.69	0.17	56.25	27.05	16.99
RBFN	36.88	72.72	27.27	63.11	60.15	10.16	-9.61	0.86	0.73	-0.61	13.05	-0.055	93.75	9.411	-3.799
RF	33.33	57.5	42.5	66.66	59.39	38.63	9.16	1.15	1.27	-0.085	38.79	0.08	64.58	27.05	8.67

V. CONCLUSION

This paper has introduced a new data classification algorithm based on the behavior of social spider called SSO algorithm. The proposed SSO algorithm are validated using a set of four dataset such as DJI dataset, three own dataset gathered from Yahoo finance on the basis of daily, weekly and yearly. The empirical result states that the proposed algorithms perform well and it is noted that the maximum classification performance is achieved by the SSO algorithm. In future, the proposed SSO algorithm can be employed to various real time classification processes.

REFERENCES

1. Qi, Yanjun, Ziv Bar-Joseph, and Judith Klein-Seetharaman. "Evaluation of different biological data and computational classification methods for use in protein interaction prediction." Proteins: Structure, Function, and Bioinformatics 63.3 (2006): 490-500.
2. S Tu, My Chau, Dongil Shin, and Dongkyoo Shin. "A comparative study of medical data classification methods based on decision tree and bagging algorithms." 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. IEEE, 2009.
3. S Schmidtler, Mauritius AR, and Roland Borrey. "Data classification methods using machine learning techniques." U.S. Patent No. 7,937,345. 3 May 2011.
4. Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K., 2018. Financial crisis prediction model using ant colony optimization. International Journal of Information Management.
5. Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. (2017). Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. Journal of King Saud University-Computer and Information Sciences.
6. Mostafa, Mohamed M. "Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait." Expert Systems with Applications 37.9 (2010): 6302-6309.
7. Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. (2017). Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. Journal of King Saud University-Computer and Information Sciences.
8. Lee, Tian-Shyug, and Chih-Chou Chiu. "Neural network forecasting of an opening cash price index." International Journal of Systems Science 33.3 (2002): 229-237.
9. Khan, Md Ashraful Islam. "Financial volatility forecasting by nonlinear support vector machine heterogeneous autoregressive model: evidence from Nikkei 225 stock index." International Journal of Economics and Finance 3.4 (2011): 138.
10. Tay, Francis EH, and Lijuan Cao. "Application of support vector machines in financial time series forecasting." Omega29.4 (2001): 309-317.
11. Deboeck, Guido. Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets. Vol. 39. John Wiley & Sons, 1994
12. Graham, B., 1949. Intelligent Investor. s.l.:Harper.
13. Barsky, R. B. & De Long, J. B., 1992. Why does the stock market fluctuate?, s.l.: National Bureau of Economic Research
14. Hendershott, Terrence, and Pamela C. Moulton. "Automation, speed, and stock market quality: The NYSE's hybrid." Journal of Financial Markets 14.4 (2011): 568-604.
15. Cuevas, E., Cienfuegos, M., Zaldívar, D. and Pérez-Cisneros, M., 2013. A swarm optimization algorithm inspired in the behavior of the social-spider. Expert Systems with Applications, 40(16), pp.6374-6384.
16. James, J.Q. and Li, V.O., 2015. A social spider algorithm for global optimization. Applied Soft Computing, 30, pp.614-627.
17. Cuevas, E., Cortés, M.A.D. and Navarro, D.A.O., 2016. Social-Spider Algorithm for Constrained Optimization. In Advances of Evolutionary Computation: Methods and Operators (pp. 175-202). Springer International Publishing.
18. Tawhid, M.A. and Ali, A.F., 2016. A simplex social spider algorithm for solving integer programming and minimax problems. Memetic Computing, 8(3), pp.169-188.
19. Nguyen, P.H., Wang, D. and Truong, T.K., 2017. A Novel Binary Social Spider Algorithm For 0-1 Knapsack Problem. International Journal of Innovative Computing Information and Control, 13(6), pp.2039-2049.
20. James, J.Q. and Li, V.O., 2016. A social spider algorithm for solving the non-convex economic load dispatch problem. Neurocomputing, 171, pp.955-965.
21. Klein, C.E., Segundo, E.H., Mariani, V.C. and Coelho, L.D.S., 2016. Modified social-spider optimization algorithm applied to electromagnetic optimization. IEEE Transactions on Magnetics, 52(3), pp.1-4.
22. Tawhid, M.A. and Ali, A.F., 2017. A hybrid social spider optimization and genetic algorithm for minimizing molecular potential energy function. Soft Computing, 21(21), pp.6499-6514.
23. Chandran, T.R., Reddy, A.V. and Janet, B., 2016, February. A social spider optimization approach for clustering text documents. In Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2016 2nd International Conference on (pp. 22-26). IEEE.
24. Zhou, Y., Zhou, Y., Luo, Q. and Abdel-Basset, M., 2017. A simplex method-based social spider optimization algorithm for clustering analysis. Engineering Applications of Artificial Intelligence, 64, pp.67-82.
25. Maurya, L., Mahapatra, P.K. and Kumar, A., 2017. A social spider optimized image fusion approach for contrast enhancement and brightness preservation. Applied Soft Computing, 52, pp.575-592.
26. Agarwal, P., Singh, R., Kumar, S. and Bhattacharya, M., 2016. Social spider algorithm employed multi-level thresholding segmentation approach. In Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2 (pp. 249-259). Springer International Publishing.

27. Lubin, T. B. (2007). The evolution of sociality in spiders. In H. J. Brockmann (Ed.), Advances in the study of behavior (Vol. 37, pp. 83–145).
28. Aviles, L. (1986). Sex-ratio bias and possible group selection in the social spider Anelosimus eximius. The American Naturalist, 128(1), 1–12.
29. Burgess, J. W. (1982). Social spacing strategies in spiders. In P. N. Rovner (Ed.), Spider communication: Mechanisms and ecological significance (pp. 317–351). Princeton, NJ: Princeton University Press.
30. Maxence, S. (2010). Social organization of the colonial spider Leucauge sp. in the Neotropics: Vertical stratification within colonies. The Journal of Arachnology, 38, 446–451.
31. Avilés, L. (1997). Causes and consequences of cooperation and permanent-sociality in spiders. In B. C. Choe (Ed.), The evolution of social behavior in insects and arachnids (pp. 476–498). Cambridge, MA: Cambridge University Press.
32. Rayor, E. C. (2011). Do social spiders cooperate in predator defense and foraging without a web? Behavioral Ecology & Sociobiology, 65(10), 1935–1945.
33. Gove, R., Hayworth, M., Chhetri, M., & Rueppell, O. (2009). Division of labour and social insect colony performance in relation to task and mating number under two alternative response threshold models. Insectes Sociaux, 56(3), 19–331
34. Jones, T., & Riechert, S. (2008). Patterns of reproductive success associated with social structure and microclimate in a spider system. Animal Behaviour, 76(6), 2011–2019.
35. Damian, O., Andrade, M., & Kasumovic, M. (2011). Dynamic population structure and the evolution of spider mating systems. Advances in Insect Physiology, 41, 65–114
36. https://archive.ics.uci.edu/ml/datasets/dow+jones+index
37. https://in.finance.yahoo.com/quote/CSV/history/