International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Optimal Feature Subset Selection for Imbalanced Class Data using SMOTE and Binary ALO Algorithm

K. Jayanthi, L. R. Sudha

Abstract: ***Feature selection in multispectral high dimensional information is a hard labour machine learning problem because of the imbalanced classes present in the data. The existing Most of the feature selection schemes in the literature ignore the problem of class imbalance by choosing the features from the classes having more instances and avoiding significant features of the classes having less instances. In this paper, SMOTE concept is exploited to produce the required samples form minority classes. Feature selection model is formulated with the objective of reducing number of features with improved classification performance. This model is based on dimensionality reduction by opt for a subset of relevant spectral, textural and spatial features while eliminating the redundant features for the purpose of improved classification performance. Binary ALO is engaged to solve the feature selection model for optimal selection of features. The proposed ALO-SVM with wrapper concept is applied to each potential solution obtained during optimization step. The working of this methodology is tested on LANDSAT multispectral image.***

Keywords: ***Feature selection, SMOTE, Binary Antlion Optimization algorithm, SVM classifier, Remote sensing.***

I. INTRODUCTION

Classification of remote sensing imagery is an important task for land cover image analysis. Classification algorithms are generally classified into pixel oriented and object oriented [1]. In pixel based classification, information processing is carried out at pixel levels without integrating the structural and spatial information. In object based classification, an image can be segmented into meaningful objects. Each object consists of group of pixels in which spectral, spatial, textural features are encapsulated. It reduces the redundant spatial details and minimizes the spectral heterogeneity of pixels [5].

Most of the documented methods for selection of optimal features did not look into the problem of imbalanced class data. The class imbalance occurs when the sample space of some classes are more whereas sample space of few other classes are less in the total sample data. This unequal distribution of classes can lead to diminish the classification performance due to the absence of important features of the minority class in the feature subset. Data sampling is often used by the researchers to balance the class distribution prior to feature selection. This problem is effectively addressed by randomly duplicate the rare samples using SMOTE method for preventing the discrimination of rare samples features in the significant feature sub set [2].

Feature Selection (FS) algorithms explore the data with the objective of eliminating or reducing the noisy, irrelevant and redundant components while simultaneously improving the classification performance. Feature selection methodologies use either filter based or wrapper based approaches for preserving the classification accuracy. Filter methods are data dependent in which less important features are removed with the use of statistical methods. Wrapper methods are classifier dependent in which each searched subset in the iterative process is given to the classifier to evaluate the classification performance. It is computationally exhaustive procedure with more number of features. To reduce the computational cost we need efficient search tool for searching best feature subset for representing the classes. Nature inspired computing algorithms are utilized for the feature selection problems in the existing literature. Some of the methodologies stuck into the local optima condition [6-8].

In this proposed methodology, binary version of ALO algorithm is implemented to hunt significant feature subset effectively without trapping into local optima condition due to its better intrusion and wandering capability in problem plane.

The structure of the paper contains six sections. Occurrence of class imbalances in high dimensional data problem is analyzed in Section-II. FS optimization model is formulated in Section-III. Section-IV explains the structure of binary ALO and its implementation of SMOTE and wrapper based proposed optimal feature subset selection methodology. Experimental results are presented and discussed in Section-V. Findings and applicability of the Binary ALO methodology are given in the concluding Section-VI.

II. CLASS IMBALANCE IN HIGH DIMENSIONAL DATA

A dataset is considered to be imbalanced if the numbers of samples are not equally distributed between classes [3]. The problem of Classification in imbalanced datasets is controlled by data resampling method. Random oversampling and undersampling are the two commonly employed resampling techniques used for balancing.

Revised Manuscript Received on January 18, 2020.
* Correspondence Author
K. Jayanthi*, Assistant Professor, Department of Computer Application, Govt. Arts College, Chidambaram, India. E-mail: jayanthirab@gmail.com
L. R. Sudha, Associate Professor**,** Department of Computer Science & Engineering, Annamalai University, Annamalainagar, India.

Random oversampling duplicates the instances of minority class in which instances are randomly selected.

The scheme of oversampling the minority classes results in the reduction of imbalance in the sample space and also yields the better classifier output.

SMOTE is an oversampling approach that balances the minority class's representation before feature selection by creating "synthetic" instances instead of oversampling with replacement. Each minority class is taken under consideration for the introduction of synthetic features in the line segments and joins all its k-nearest neighbours. Artificial copies of the samples are in the range of 100% to 500% of minority class instances. Oversampling of the minority class sample depends on the required quantity of oversampling. This method randomly chooses the neighbours from the k nearest neighbours [2].

N is the number samples in the class which need to be balanced, Oversampling needed in percentage is A and k is the number of nearest neighbours. Number of total minority class samples after balancing is given by

$$NB_{MCS} = \frac{A}{100} * N \quad (1)$$

Synthetic samples are generated as follows

1. The difference between the feature samples in the minority class under consideration and its nearest neighbors are evaluated.
2. Multiply the difference value by a generated random number between 0 and 1.
3. Augment these synthetic feature samples to the class under attention.

III. FEATURE SELECTION OPTIMIZATION MODEL

Feature selection task addresses the difficulty of handling large number features by searching only an opt subset with relevant attributes from the extracted attribute variables. This based on the principle that few significant features are enough to charaterise the object. This mechanism improves the classification task by eliminating redundant and irrelevant features. FS process is modeled with the twin objective of increasing the accuracy in class identification and reduction in number of attributes, thereby reducing the difficulty in computation. Classification objective is coined using Kappa coefficient and fraction of number of features in the subset to total number features is used as second term in optimization model [4, 10].

$$Min\ Z = W\ (1 - Kappa) + (1 - W)\left(\frac{N_{SF}}{N_{TF}}\right) \quad (2)$$

Where W is the weightage factor in the range of (0, 1), NSF represents no. of sub-set features and NTF represents no. of total features.

Kappa represents the classification metric which is given by the relation

$$Kappa = \frac{Total\,Accuracy - Random\,Accuracy}{1 - Random\,Accuracy} \quad (3)$$

Total accuracy is the ratio of true classification samples to the total samples which is given by

$$Total\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Random accuracy is function of actual and predicated sample classes, and square of total samples.

*Random Accuracy= (Actual false*Predicted False + Actual True*Predicted True) / (Total*Total)*

Feature selection model requires efficient algorithm to search the optimal subset in the multidimensional problem search space. Recently Ant Lion Optimization (ALO) algorithm based on mimicking the food seizing mechanism of antlions and their food ants in their habitat has been developed for solving complex optimization objective. Its capability to solve the variety of real world problems is tested by applying ALO on various problem domains. Promising results prove its adaptability for problem solving in multi-dimensional search space of real world applications. Due to its inherent exploration and exploitation capability, this methodology utilizes its mechanism for solving the formulated feature selection model. Decision variables are based on whether particular feature is part of the subset or not. This makes binary representation of variables is necessary for the feature selection model.

IV. ALO FOR FEATURE SELECTION IN BALANCED CLASS DATA

A. Overview of Binary Ant Lion Optimization algorithm

In this system, preys (ants) stride the habitat plane and the predators (antlions) create pits in the habitat to engulf the insect prey for food. Few strategic steps followed by ant and antlion for the optimal solving procedure are random movement of ants, creating the traps, engulfment of ants in the created pit, prey gathering and pits rebuilding. ALO imitates this mechanism for searching the optimal solution in multi-dimensional search space.

The comprehensive ideal solution to the problems are effectively obtained by ALO algorithm due to better travelling across the problem plane performed as unintended movement & arbitrary selection of ants and antlions. The local optima are highly prevented during the unintended movement and arbitrary selection of antlions based on their health. During optimization, antlions take the position of healthy ants and this location is saved. The healthier antlion in that generation is stored, named Elite.

Step 1: Initialization

ALO Parameters such as total ants & antlions, maximum planned generation, stopping criteria are initialized. The initial population of the search agents; ants and antlions are generated randomly asolution of optimal feature subset selection model requires binary representation of the control variable which represents a particular feature. Significant features are represented by 1 and insignificant features are by 0. ALO belongs to the group of natural process mimicking optimization procedure which replicates the capturing of food process in Doodlebugs insects (larve of antlions). ALO algorithm for binary space is described in the following section [9].

$X_i^0 = [x_{i,1}^0, x_{i,2}^0, \ldots, x_{i,k}^0, \ldots, x_{i,n}^0]$
(4)

$$x_{i,k}^0 = \text{round}\ (rand\,(0,1]\,)$$
(5)

X_i^0 is the initial position of search agent i. $x_{i,k}^0$ is the binary vector in k^{th} dimension of the earch agent i.

Step 2: Evaluating fitness value

Evaluating the health of the searchers, a fitness value is found using the objective to be achieved. The antlion with better health from the initial generation can be stored as elite.

Step 3: Trapping in antlions' pits

The unintended movements of ants in the problem plane are affected via Ant lions' created pits. It directs the prey insects towards unknown search regions. This can be observed by the following equations

$$C_k^t = Antlion_j^t + C^t \quad (6)$$

$$D_k^t = Antlion_j^t + D^t \quad (7)$$

Where C_k^t - the least value of 'k' in't', D_k^t - highest value of 'k' in 't', C^t – the least of all variables in 't' and D^t is the highest value of all 'k' in t, k denotes variables and t denotes iteration.

Step 4: Sliding ant in the direction of antlion

During the process of building traps, traps are built based on antlions' strength and ants are necessary to shift their position randomly. The roulette wheel is used to assign antlion depends on the fitness or simply health. When they found an ant in the trap, they started throwing sand from the middle to over the pit until the ant tumble into the created pit. This step is modelled as the range of ant's unintended circular movement that changed adaptively depending on present iteration level.

$$C^t = \frac{C^t}{I};\ D^t = \frac{D^t}{I};\ \ I = f(w,t,It) = 10^w \frac{t}{It} \quad (8)$$

Step 5: Normalize the Random walks of ants

Ant's movement is random in nature during the food searching process. This unintended movement may be imitated to the next positional change as

X(t) = [0,cums(2s(t_1) – 1),cums(2s(t_2) –1, ….,cums(2s(t_(N_I)) – 1)]
(9)

where cums calculates the cumulative sum and r(t) is defined as follows:

$$s(t) = \begin{cases} 1 \ if \ \ rand > 0.5 \\ 0 \ if \ \ rand \le 0.5 \end{cases} \quad (10)$$

To retain the unintended movement within the problem plane, normalization is carried based on Equ. (9) is

$$X_{i,k}^t = rand * (D_k^t - C_k^t) + C_k$$

(11) Where $X_{i,k}^t$ - kth variable position in that generation of a particular ant i.

Step 6: Catching ants for food and recreate the pit

Every ants' health in its present position is found using objective selected. Modify antlion's location to the ant's present location if the hunted ant has a better health otherwise keep the original antlion position for the next iteration.

$$Antlion_j^t = Ant_i^t \quad if\ f(Ant_i^t) > f(Antlion_j^t)$$

(12) Step 7: Elitism

Healthiest antlion acquired to be preserved, named elite. Superior elite antlion can able to disturb the ants location thereby updating of ants position in the problem plane. Ants should perform unintended movements based on the antlions fixed and elite. It is performed with crossover & mutation operators as

$$Ant_i^t = crossover\ (\ Antlion_a^t\ ,\ Antlion_e^t\ )$$

$$Ant_i^t = \begin{cases} Antlion_a^t & if\ rand \le CRP \\ Antlion_e^t & otherwise \end{cases}$$

(13)

Where CRP - Probability of crossover value lies between (0, 1).

In soft computing, mutation is considered as alteration with randomness. Mutation is influenced by ant vector, mutation probability defined & numeral generated. Mutation (Binary) with $Ant_{i,k}^t$ after crossover is

$$Ant_{i,k}^t = \begin{cases} 1 - Ant_{i,k}^t & if\ rand \le MP \\ Ant_{i,k}^t & otherwise \end{cases} \quad (14)$$

Where $Ant_{i,k}^t$ in the left side of the above equation represents the final ant position after mutation operation and right side term represents ant position after crossover operation between antlion 'a' and elite antlion 'e'. MP - Probability of mutation value within 0 to 1.

Step 8: Convergence

If the convergence measures are not satisfied then the ants update their position for further exploration of the searching space otherwise the searching procedure is terminated. The position of elite antlion gives the best possible result. Convergence criteria may be either an acceptable solution found or no betterment in health is possible or a chosen number of unintended movement finished.

B. Binary ALO implementation in Feature Selection

Feature extraction task extracts region attributes by gathering relevant information regarding objects in the image. This process reduces the computational difficulty associated with the image analysis by extracting only relevant features. Computational resources required and complexities involved in the image analysis are further reduced in the FS process. It overcomes the dimensionality problem by identifying the optimal sub set of features that can effectively describe the image objects without sacrificing classification accuracy. In the proposed methodology, the multi-spectral image is segmented into objects and features are extracted. The features picked are used to train the SVM classifier. Instances of imbalanced low samples classes sample are randomly generated based on SMOTE methodology for balancing the minority classes. The dimension of Binary search agent is equal total attributes extracted for the objects. Decision variables in a particular dimension is one means that subset considers that a particular feature as a significant feature. Fitness of the each searched sub set is evaluated based on feature selection objective.

Steps involved in the implementation of Binary ALO for optimal FS is presented in fig. 1.

.

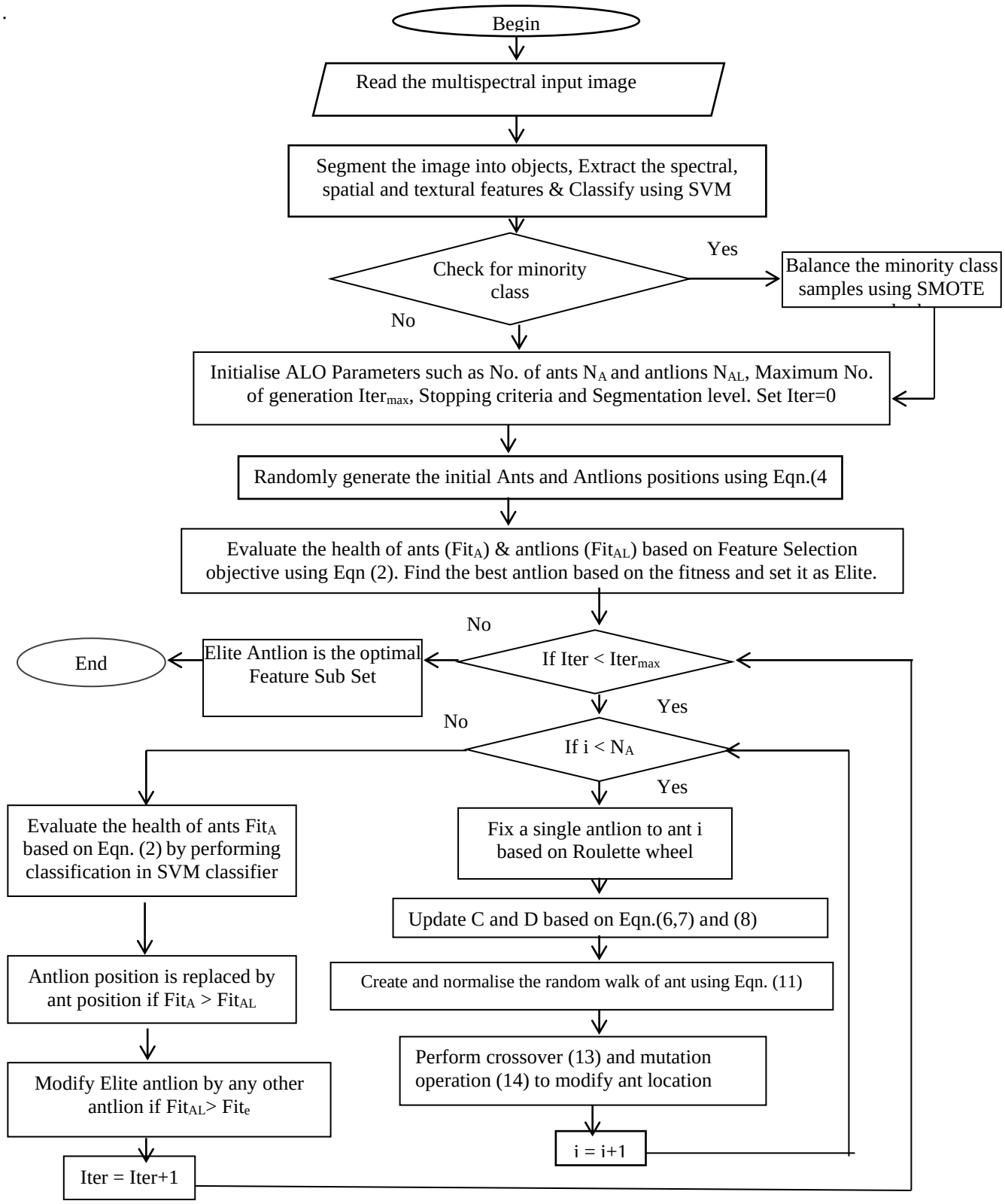

Fig.1: Flowchart of Binary ALO Implementation for Optimal Feature Selection

V. EXPERIMENTAL RESULTS AND DISCUSSION

Binary ALO based feature selection with SMOTE balancing methodology is tested on the land cover data set attained from Landsat-7 imagery. The area selected for classification is the Tiruchirappalli city in Tamil Nadu. It is located at "100 39'N780 33'E100 47'N 780 51'N" covering the area of 540 sq.km with 30 meter resolution.

This image is segmented into objects and classified into six classes namely agriculture, forest, buildup urban, buildup rural, water bodies and barren land based on SVM classifier. The original image (fig.1) and classified image (fig.2) are given below.

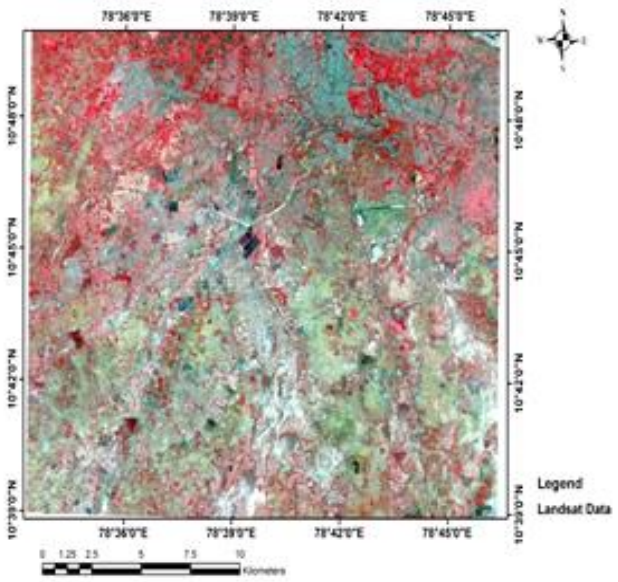

Fig. 2 : Landsat-7 image (False Colour Composite)

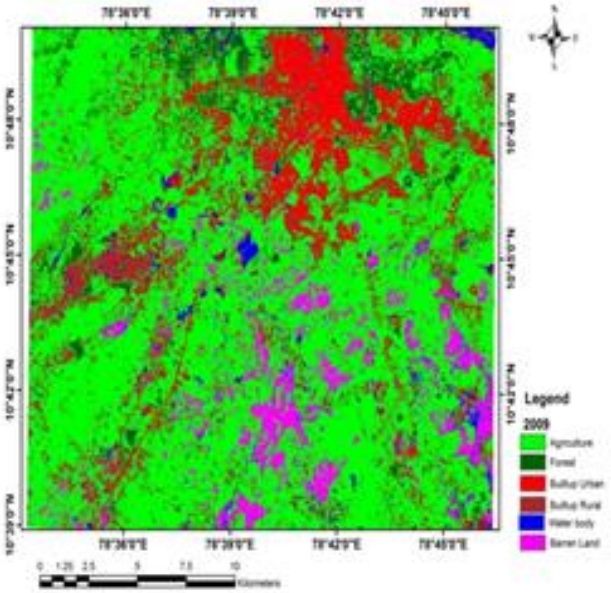

Fig. 3: Classified image

The selected SVM has proven track record in the classification of high dimensional remote sensing data. The feature set of an object consists of seventy spectral, textural, and spatial attributes. Test set contains randomly chosen 2580 samples to compare the performance of complete feature set and significant feature set.

The random sample selection results least number of samples for water body class. Sample instances are randomly added in test set based on SMOTE concept for achieving balance in the classification task. Optimal parameter value is tuned for SVM using '10-fold cross validation' method. Consistency is achieved by centering and rescaling of variables before classification. This validation model is applied to complete feature set. The confusion matrix for test data having 70 features is exhibited in table-1.

Binary ALO parameters fixed to solve the FS model are ants: 30, antlions: 20, iterations: 100, probability of crossover: 0.70, rate of mutation: 0.05. ALO achieves maximum objective value with 24 significant features. With Significant feature subset, SVM realizes the overall accuracy of 89.849% and Kappa value of 0.832. SVM results with optimum sub set are tabulated in Table-2.

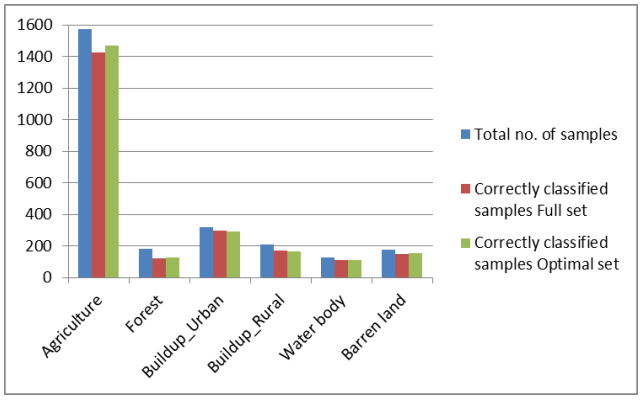

Fig.4. Performance metric analysis of test dataset - SVM classifier

Table 1: Confusion matrix – Accuracy in class prediction (With Complete 70-features)

	No. of samples	Agriculture	Forest	Buildup Urban	Buildup Rural	Water body	Barren land	Producer's Accuracy (%)
Agriculture	1571	1426	34	14	53	11	33	90.71
Forest	181	26	121	2	8	4	20	66.85
Buildup Urban	316	3	1	295	13	1	3	93.35
Buildup Rural	207	8	3	23	171	0	2	82.61
Water body	129	8	9	0	3	108	1	83.72
Barren land	176	11	4	7	6	0	148	84.09
User's Accuracy (%)		96.22	70.35	8.51	67.32	87.10	71.50	

Overall Accuracy: 87.946% Kappa: 0.804

Table 2: Confusion matrix – Accuracy in class prediction (With 24 significant features)

	No. of samples	Agriculture	Forest	Buildup Urban	Buildup Rural	Water body	Barren land	Producer's Accuracy (%)
Agriculture	1571	1467	12	16	43	10	23	93.38
Forest	181	25	129	1	6	3	17	71.27
Buildup Urban	316	3	1	290	18	0	4	91.77
Buildup Rural	207	9	3	31	164	0	0	79.23
Water body	129	8	9	0	0	112	0	86.82
Barren land	176	8	4	5	3	0	156	88.70
User's Accuracy (%)		96.51	81.65	84.55	70.09	89.60	78.11	

Overall Accuracy: 89.849% Kappa: 0.832

It is found from Tables 1 & 2 and fig.4, the optimal feature subset resulted from the proposed methodology achieves better classification accuracy of 0.8985 and kappa coefficient of 0.832 with 24 features. But the original imbalanced dataset having 70 features obtained the classification accuracy of only 0.8795 and kappa coefficient 0f 0.8040. This proves the adoptability of the proposed method for feature selection.

VI. CONCLUSION

Binary ALO-SVM method aimed at feature selection to address the dimensionality problem has been proposed. The objectives in FS are improving the classification performance with reduction in features required for representing the objects. In land cover analysis, minority classes are found due to the irregularities in bio-physical components. The imbalanced samples of minority class are balanced using SMOTE method. SVM classifiers are utilized to classify the image objects during optimal subset selection mechanism as well as with full feature data set. ALO-SVM method is applied to sample image taken from Landsat-7 for validation. The converged results are promising with improvement in classification accuracy and converging to less number optimal features for representing the object classes. Comparisons are provided with original data feature performance measures and feature selected sub set performance measures.

REFERENCES

1. Archibald, R., Fann, G., "Feature Selection and Classification of Hyperspectral; Images with Support Vector Machines", IEEE Geoscience and Remote Sensing Letters, Vol. 4 (4), pp.674–677, 2007.
2. Chawla, N.V., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P., "SMOTE: synthetic minority oversampling technique", Journal of Artificial Intelligence Research, Vol. 16, pp.321–357, 2002.
3. He, H., Garcia, E. A., "Learning Form Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 21(9), pp.1263–1284, 2009.
4. Huang Ch. L., Wang Ch. J., " A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications, Vol. 31, pp.231-240, 2006.
5. Ma, L., Li, M., Gao, Y., Chen, T., Ma, X., Qu, L., "A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation", IEEE Geosci. Remote Sens. Lett., Vol. 14(3), pp.409–413, 2017.
6. Maldonado. S., Weber. R., " Feature selection for high-dimensional class-imbalanced datasets using Support Vector Machines Machines", Inf. Sci., Vol. 186, pp.228–246, 2014.
7. Mirjalili, S., Lewis, A., "S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization. Swarm and Evolutionary Computation", Vol. 9, pp.1-14, 2013.
8. Senthilnath, J., Kulkarni, S., Benediktsson, J.A., Yang, X.S., "A novel approach for multispectral satellite image classification based on the bat algorithm", IEEE Geosci. Remote Sens. Lett., Vol. 13(4), pp.599–603, 2016.
9. Seyedali Mirjalili, "The Ant Lion Optimizer", Advances in Engineering Software, Vol. 83, pp.80-98, 2015.
10. Yang, H., Du, Q., Chen, G., "Particle swarm optimization-based hyper-spectral dimensionality reduction for urban land cover classification", IEEE Journal of Selected Topics in Application of Earth Observation & Remote Sensing., Vol. 5(2), pp. 544–554, 2012.

AUTHORS PROFILE

K. Jayanthi, working as Assistant Professor in the Department of computer application, Government Arts College, Chidambaram, India. She completed M.C.A from Bharathidhasan University and M.Phil [Computer science] in Annamalai University. She is pursuing Ph.D in Annamalai University. She is doing research in the area of image processing application to remote sensing. She has published 4 research articles in international journals and presented research papers in 6 international / national conferences.

Dr. L. R. Sudha, working as Associate Professor in the Department of computer Science & Engineering, Annamalai University, Chidambaram, India. She completed M.E and Ph.D in Computer Science & Engineering from Annamalai University during 2006 & 2014 receptively. She has 20 years of teaching and 16 years of research experience. In her credit, she published 15 research articles in international journals and presented papers in 20 international/national conferences. Her area of interest includes image processing and machine learning.