An Ant Colony Optimization Based Feature Selection for Data Classification

Rajesh Dwivedi, Rahul Kumar, Ebenezer Jangam, Vishnu Kumar

***ABSTRACT---** Feature selection is important process in the task of classification and clustering when the large number of feature gets extracted. In feature selection for n number of feature there are 2^n feature subsets means every feature have two possibilities first possible is that particular feature would be selected for classification and other is would not be selected for classification. So finding a relevant feature subset in appropriate time is a NP-Hard problem. To avoid this problem, the approximation algorithm is used that gives the near optimal solution are four types including filter, wrapper, embedded and hybrid techniques. Many of the swarm intelligent algorithms that simulate the social behaviour of living beings are used as feature selection algorithms. The proposed method using the one of the swarm intelligent algorithm for feature selection based on ant colony optimization. This algorithm is combined with the Support vector machine classifierfor selecting the more appropriate and useful features.*

***Keywords:** Feature Selection, Ant Colony Optimization, Data Classification*

1. INTRODUCTION

In machine learning, feature selection is the way toward choosing a subset of relevant features for development of a predictive model. The main hypothesis behind feature selection is that data contains many features that are inappropriate or irrelevant and redundant thus to be eliminated without suffering much loss of information. Removal of these features is important because it unnecessary makes the model complex; it takes larger training time and reduces accuracy of model. Relevant and redundant are two distinct terms. A relevant feature can also be redundant if it strongly depends on some other relevant features.

Feature selection is carried out with combination of searching technique, that makes the feature subsets and a performance evaluator that evaluates the performances of feature subsets. The process of feature selection can be done by combining two processes named as searching process and selection process. Therefore the feature selection method can be categories in three types of searches named as Complete Search, Stochastic Search and Heuristic Search.

In feature selection process, Apart from the diminishment in the features counts, Accuracy of feature subsets is fundamentally vital. That's why in this work the feature selection method is combined with the classifier algorithm to learn and model the underlying processes. There are various techniques where the feature selection process is combined with classification techniques. The three basic techniques for feature selection are given as follows:-

1.1 Filter Based feature Selection Method

In the filter based feature selection method [1-5], a filtering process is used before the classification task, that's why selected features are independent from the classification algorithm [6]. In this method features are selected by considering proxy measure instead of using the error rate to rank a feature subset. The proxy measure contains the point wise mutual information, Pearson product-moment correlation coefficient etc. filter based method less computationally complex that wrapper methods because it does not evaluate the feature subset by using classification algorithms. The selected features using this approach are more generalized features but give the lesser prediction accuracy than wrapper method. The procedure for feature selection using filter approach presented in fig. 1.

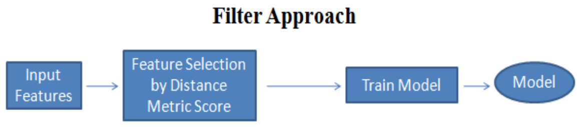

Fig. 1: Filter based feature selection approach

1.2. Wrapper Based feature Selection Method

In this wrapper approach [7-8], the selection of feature process is combined with the classification method used, so it does not selects the general feature subset which is independent from the classification algorithm. This method has more computational complexity than the other feature selection approaches because it uses classification accuracy to rank a feature subset. The performance of prediction in wrapper based approach is better than other models by considering the specific classification technique (which is used in the learning process). The procedure for feature selection using wrapper approach presented in fig. 2.

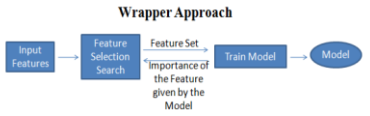

Fig. 2: Wrapper based approach for feature selection

Revised Manuscript Received on February 11 , 2019.

Rajesh Dwivedi, Department of Computer Science and Engineering, Vignan Foundation For Science Technology and Research Guntur, A.P, India. (E-mail: anubhav.dwivedi8@gmail.com)

Rahul Kumar, Department of Computer Science and Engineering, Vignan Foundation For Science Technology and Research Guntur, A.P, India.

Ebenezer Jangam, Department of Computer Science and Engineering, Vignan Foundation For Science Technology and Research Guntur, A.P, India.

Vishnu Kumar, Department of Computer Science and Engineering, Vignan Foundation For Science Technology and Research Guntur, A.P, India.

1.3. Embedded Feature Selection Method

The embedded feature selection technique performs feature selection as a major aspect of the learning strategy and is normally particular to given learning machines. These are the recently proposed methods, In this method feature selection the learning technique take the benefit of its own feature selection and perform the task of feature selection and classification simultaneously. Example for this approach is construction of a linear model by LASSO method, having a regression coefficients penalty of L1 in which many of them are shrinking to zero. The procedure for feature selection using wrapper approach presented in fig.3.

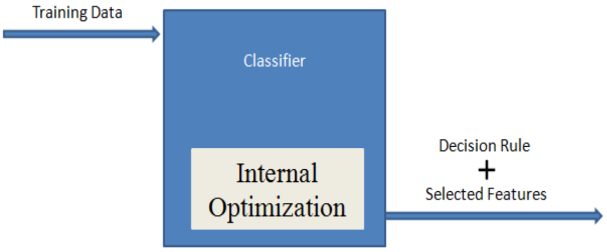

Fig.3: Embedded based feature selection approach

1.4. Hybrid Feature Selection Approach

Hybrid based feature selection is the mixture of the wrapper and filter approaches as shown in fig. 4.

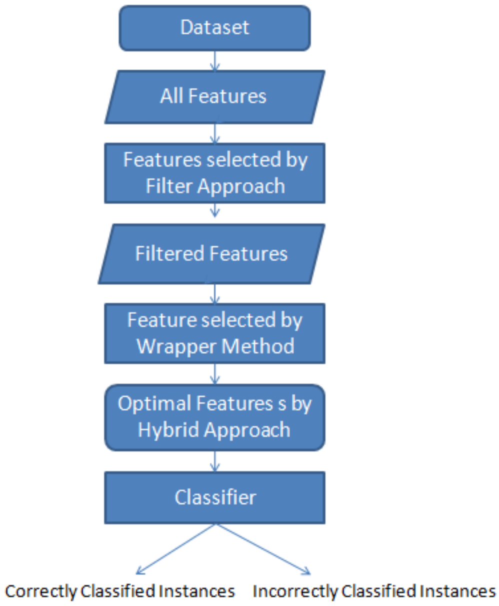
Dataset
All Features
Features selected by Filter Approach
Filtered Features
Feature selected by Wrapper Method
Optimal Features s by Hybrid Approach
Classifier
Correctly Classified Instances
Incorrectly Classified Instances

Fig. 4: Hybrid model for feature selection

Many of the optimization approaches that simulate the behaviour of living beings such as Ant colony optimization, Particle Swarm optimization, Artificial bee colony optimization, bat opimizaion are used as feature selection algorithm.

2. RELATED WORKS

Song.M.H.et al. [9] developed a system for arrhythmia classification. They used Linear Discriminant Analysis (LDA) combined with SVM classifier for the dimensionality reduction. In their work total numbers of seventeen features are getting extracted by using the wavelet transform. By using LDA only four features are getting selected. The performance of the selected feature subset is compared with the all features and features selected by Principle component analysis (PCA) which results the selected feature subset gives better performance among these. In this work the SVM is also compared with the Multi-layer perceptron (MLP) and the fuzzy inference System (FIS) which results that SVM classifier gave better performance among these.

Lin S.W. et al. [10] developed a system that uses simulated annealing approach for the feature selection in SA-SVM method. The SVM classifier has the main difficulty for setup of the parameter values for the kernel function. If the values of parameter is not properly set then the SVM classifier gives the wrong outcomes. In the wrapper based approach classification technique is the parameter for choosing the best subset. The authors also compared their results with the original set of features and found the feature selection gives the better prediction accuracy. The outcomes also conclude that SA-SVM approach give advantage for setting the parameters values.

Gutlein. M. et al. [11] discussed in their work that linear forward selection techniques helps for reducing the feature expansion in every forward selection steps. Which inferred that this approach is faster and find the smaller subsets, and can also increase the accuracy of the forward selection. These techniques also compared with complete forward selection in terms of computational cost, time and overfitting which results that linear forward selection takes less computational time and less overfitting.

Esseghir M. A [12] developed a feature selection method based on the greedy randomized adaptive search procedure (GRASP). It is hybrid based approach that uses the combination of wrapper and filter based approaches. They tested their approach with the five machine learning datasets and found that it select the relevant and the important features.

Alper U et al. [13] discussed in their work for a hybrid feature selection approach named as maximum relevance and redundancy PSO (MRRPSO) that uses the combination of filter and wrapper method. The algorithm is based on the Particle Swarm Optimization (PSO). Filter method is depends on the mutual information and gives the relevant feature subset without considering the classification algorithm. The designed model is also compared with hybrid feature selection algorithm based on genetic algorithm and a wrapper method using PSO which results that this MRRPSO is competitive in terms of accuracy and time complexity.

Swati S. et al. [14] developed a feature selection approach that uses filtering based on Signal to noise ratio score (SNR) and Particle swarm optimization for the high dimensional microarray dataset. They used the K-Means and SNR score for grouping the data and to score the genes respectively. The genes having high scores are selected from each cluster to create a new feature subset. After that the subset is passed to PSO approach that gives the optimized feature subset. For the classification task the probabilistic neural network, K-nearest neighbor and support vector machine are used. The model is also compared with many feature selection approaches and found that it gives the better feature subset than others.

Thananan prasartvit et al. [15] discussed in their work that how the artificial bee colony algorithm is used for selecting features from the high dimensional data. It is used for the selecting the best and relevant features using the ABC. This work uses the wrapper based approach for the feature subset evaluation that’s why the K-nearest neighbor classification technique is used along with the ABC algorithm.

Nahar J.T.I et al [16] developed a system that uses Medical knowledge for the feature selection along with computational intelligence. In this work MFS is combined with computerized feature selection process (CFS) for the feature selection and found that these algorithm gives best result with the naïve bayes classifier and the sequential minimal optimization (SMO).for the development Waikato environment is used along with the weka tool. Testing was performed by using various machine learning datasets having heart related disorders. For the accuracy testing they considered the four measures that are F-measure, true positive rate, Time and accuracy.

In the proposed work feature selection is done by using ant colony algorithm.

3. PROPOSED WORK

Ant Colony Optimization (ACO) calculation depends on the attractive social conduct of ants in looking for sustenance when ants discover a food source they leave a smelly material called as pheromone which is used to mark the path between the source and food. At the point when an Ant searching for the food it smells the odorous material and follow the path, which is having more odorous material. This ant will likewise leave pheromone on this way which builds the way’s quality, which is used to attract the other ants likewise to select the same path.

If an ant requires picking among various ways, it inclines toward the path, where the quantity of pheromone is high, which shows that more number of ants has gone via that way. Since the ants select shorter ways to take the nourishment o their settlements, the short length routes get highest level of pheromone compared to long length routes. If a way is not navigated by any subterranean insect, at that point the pheromone dissipates over time, subsequently diminishing the pheromone level of that way. In different words, diminish in pheromone powers that force the ants to investigate new ways for sustenance.

Based upon the perspective, ACO algorithms may have a place with various classes of approximate algorithms. As per the AI aspect ACO calculations are a standout amongst the best strands of swarm intelligence. The aim of swarm intelligent to analyse the behaviour of living beings that how they are interacting with each other and with the environment for example analysing the behaviour of ants, birds, honey bees, wasps, and other animal community for example, groups of winged animals or fish schools. Cases of “swarm intelligent” calculations other than ACO are those for classification and data mining motivated by living beings are artificial bee colony, particle swarm optimization.

3.1. ACO Algorithm for Feature Selection

The Simulation model is expressed by a undirected graph G= (V, E) which is completely connected. Where V is the group of vertices as $v_1,v_2,v_3,...v_n$ and E is the group of edges. Every vertex in the graph corresponds to a feature $f_1,f_2,f_3....f_n$ and edge E denotes edges linking any two vertex. In this approach every feature is represented by a vertex so count of features is equal to the count of vertices in the graph. If the graph having n vertices then $f_n=v_n$. The count of ants also equal to the no of features because it explores the search space and avoid the problem to getting trapped in local optimum. In the proposed work the count of ants equal to the count of feature subsets because every ant chooses a subset of feature. If the count of feature subsets showed by M then M = M_{ant}. Where the count of ants is denoted by M_{ant} and feature subsets are represented by F_i (which is i^{th} feature subset).The max count of features in a subset of feature is showed by $n_{maximum}$. So $0<n_{maximum}< n$ where n is the number of features in this work. Every ant starts visit from a vertex and continue visits an arrangement of vertices, this arrangement is denoted by a subset F that contain the vertices (features) f traversed by that ant. The Pheromone esteem (τ) is associated with each feature and is set to a steady at first. In the feature subset selection of $n_{maximum}$ features is carried out in five steps is as follows. The flow diagram for these steps is given in Fig 5.

Step 1:- this step comprises of M subsets, each subset having $n_{maximum}$ number of features and these M subsets are evaluated by SVM. The subset which is giving highest accuracy is known as globalbest related to leader ant and denoted by ‘$global_{best}$’.

Step 2:-In this step, every subset gets n_{filter} features by choosing a degree of features randomly, where $n_{filter}= n_{maximum}-n_{remain}$ and $n_{remain} = n_{abitrary} + n_{leader}$.

Step 3:- In the third step the $n_{arbitrary}$ features are picked, which are not yet included into partially built subset which are having highest level of pheromone and low value of cosine similarity.

Step 4:- In fourth step, the remaining n_{leader} ($n_{leader} =n_{remain}- n_{arbitrary}$) features are selected using tandem run approach. In this approach the unique features of leader subset are included to the subset under consideration having highest pheromone and low cosine similarity.

Step 5:-In this step all subsets are evaluated by using SVM and subset, which is having maximum accuracy, is marked as ‘ l_{best} ’ local best subset. If the accuracy of local best is superior then the global best then local best subset is set as global best subset and considered as leader subset for future iterations.

In this algorithm the first step execute exactly once and the remaining Step from two to five are executed until the stopping criteria not fulfilled after execution of all iterations the leader subset is considered as best subset yielding maximum accuracy.

3.2. Updation of Pheromone

Step1:-Assign beginning pheromone to all features by equation 1.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

Pheromone (f_i) = 1/ total number of features. (1)

Step2:-whenever a feature getting selected in a subset then its pheromone Updation happens by equation 2 ,3 and 4.

Fitness (f_i) = Accuracy of F_i / maximum number of feature in F_i ($n_{maximum}$) (2)

New Pheromone (f_i) = pheromone (f_i) +fitness (f_i) (3)

Pheromone (f_i) = new pheromone (f_i) (4)

3.3 Computation of Cosine similarity

Cosine similarity [17] computes the similarity between two vectors using equation 5. Suppose X and Y are two vectors having 'n' components and Xi and Yi are the i^{th} component of both the vector respectively.

$$cosine\ similarity = \frac{\sum_{i=1}^{n} XiYi}{\sqrt{\sum_{i=1}^{n} Xi^2}\sqrt{\sum_{i=1}^{n} Yi^2}} \qquad (5)$$

Arbitrariness facilitates exploration feature space which avoids the ants from choosing a similar arrangement of features. Cosine similarity prevents from selection of approximately identical features.

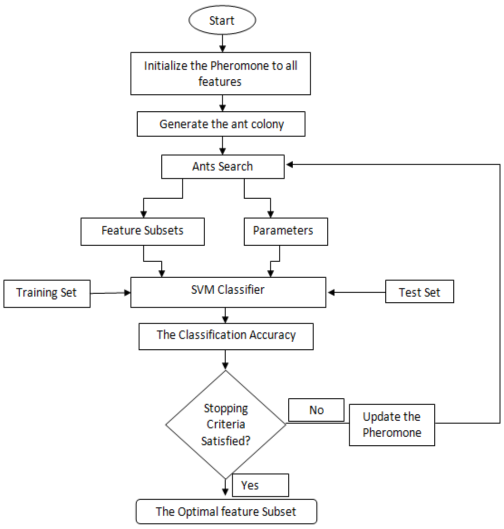

Fig. 5: Block diagram representing ACO for feature selection

3.4 Parameter Settings for ACO

The algorithm for Feature selection is carried out by having different values of $n_{maximum.}$ Setting of parameters for this work is given as follows-

$n_{maximum}$ = depends on the counts of feature to be chosen.

n_{leader} = 20% of $n_{maximum}$, Maximum number of Iteration =30.

4. RESULTS

4.1 Dataset Details

The proposed work applied on two classification dataset named as Ionosphere and Dermatology [18]. The Ionosphere dataset is having 34 attributes (Features) and 351 instances and 2 classes (binary classification) where the values of attributes can be real or integer. Whether in Dermatology dataset is having 33 attributes with 366 instances and 6 classes. In this dataset attribute values are only integer. Both the dataset can easily downloaded from UCI machine learning repository.

4.2 Classification Performance using SVM Classifier

The SVM classifier [19] plays a important and wide role in the classification because of its high accuracy and capability to deal with data of high dimension. The simply form of the classification is the binary classification that is used for separating two types of objects, one belonging from positive class (+1) and another one belonging from negative class (-1). Support Vector Machine uses two types concepts two distinguish between two classes. first one is Separation from margin and second is Kernel function. The simple two dimensional data can be classified by using a straight line. The points fall above the line belongs to one class and the points fall below the line belong from the another class.

The high dimensional data can be classified by using the hyper planes. But in binary classification multiple planes can be drawn that separates the data into two classes, so which plane will be selected for the classification? In this case the hyperplane gives maximum margin will be selected for classification. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized.

Classification of the data with best margin hyperplane is shown in fig. 6 presented below-

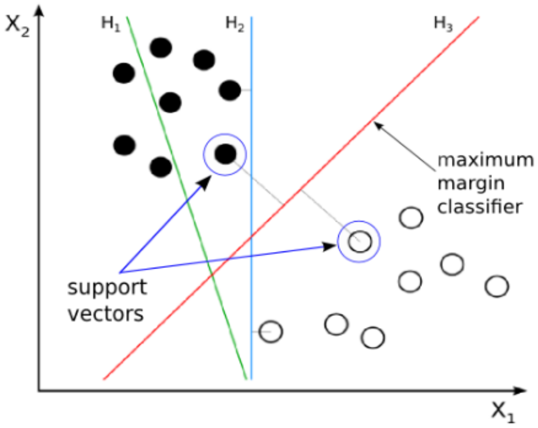

Fig. 6: Classifications of data using various hyperplane

In the fig. 6 there are two types of data points, one that are filled dots and another one that are unfilled dots and there are three planes exists named as H1, H2 and H3. The H1 does not successfully classify the data points. Plane H2 an H3 both are capable to classify data points but the H2 gives the fewer margin than the plane H3. That's why the plane H3 is selected for the classification.

Sometimes the data is not classified by hyperplanes because of its distribution in wide space so in that case we use the nonlinear separation for the classification. The SVM classifier can efficiently perform this nonlinear classification by using kernel functions. The nonlinear classification is presented in fig. 7. In figure 7 there are two types of objects,

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

one that are blue colored and another one have red colour. The objects represented in this figure cannot be separated using a linear hyperplane; Support vector machine performs this task by using kernel functions. The kernel function separates the data in the feature space by using a linear hyper plane. Some basic kernel functions are Linear Kernel, Polynomial kernel, Radial basis function kernel and Sigmoid Kernel.

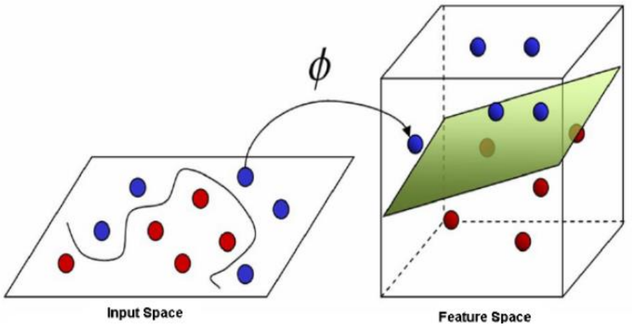

Fig. 7: Use of kernel function in SVM classifier

In this work Classification is done by using SVM Classifier in which both datasets divided into 70%-30% ratio. Where 70% of data issued for training and 30% of data used for testing. Accuracy, specificity, precision and sensitivity/recall of proposed work is calculated using formula (6), (7), (8) and (9) respectively presented below: -

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \qquad (6)$$

$$Specificity = \frac{tn}{tn+tp} \qquad (7)$$

$$Precisian = \frac{tp}{tp+fp} \qquad (8)$$

$$\frac{Sensitivity}{Recall} = \frac{tp}{tp+fn} \qquad (9)$$

Where tp and tn are the Counts of true positives and true negatives recognized by the system respectively. fp and fn are the counts of true negatives identified as positives and true positives identified as negative identified by the system respectively.

The feature selection based on ACO performed on both the dataset by selecting different number of features. Classification performance on Ionosphere dataset is described in Table 1.

Table 1: Classification Performance on Ionosphere Dataset

No. Of Feature Selected	Value of n_{filter}	Value of n_{leader}	Value of $n_{arbitrary}$	Classification Accuracy (%)
19	10	4	5	87.77
21	12	4	5	90.42
23	13	5	5	93.67
25	15	5	5	95.90
27	16	6	5	94.37

It can be observed from table 1 that classification accuracy is highest 95.90% when 25 features are getting selected. The Classification accuracy of 91.20% observed by considering all 34 features for the classification which is 4.70% less than the classification accuracy after applying feature selection using ACO. The Comparison graph on classification accuracies for different number of features is given fig 8.

Classification performance on Dermatology dataset is described in Table 2.

It can be observed from table 2 that classification accuracy is highest 91.05% when 23 features are getting selected. The Classification accuracy of 84.27% observed by considering all 33 features for the classification which is 6.78% less than the classification accuracy after applying feature selection using ACO The Comparison graph on classification accuracies for different number of features is given fig 9.

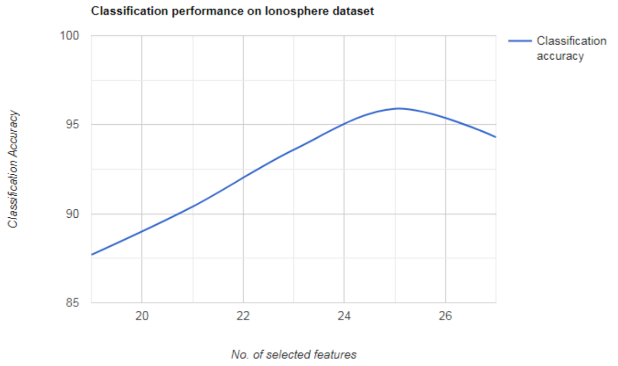

Fig. 8: Classification Performance graph for Ionosphere Dataset

Table 2: Classification Performance on Dermatology Dataset

No. Of feature Selected	Value of nfilter	Value of nleader	Value of narbitrary	Classification Accuracy (%)
19	10	4	5	83.15
21	12	4	5	85.23
23	13	5	5	91.05
25	15	5	5	89.41
27	16	6	5	87.77

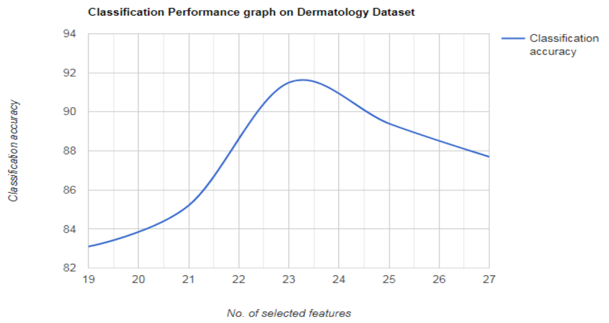

Fig. 9: Classification performance graph for Dermatology Dataset

5. CONCLUSION

In the proposed method a feature selection algorithm based on ant Colony optimization (ACO) is presented that removes the unimportant, irrelevant and redundant features and selects the more appropriate features from data having large number of features. the feature selection work can be extended by considering other bio inspired algorithm for feature selection.

6. ACKNOWLEDGEMENT

The author’s are very thankful to Dr. A.C.S Rao and Mayur wankhade for his valuable suggestions.

REFERENCES

1. M. Dash, K. Choi, P. Scheuermann, and H. Liu, “Feature selection for clustering-a filter solution,” in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, 2002, pp. 115–122.
2. C. M. Lewandowski, N. Co-investigator, and C. M. Lewandowski, “Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning,” *Eff. Br. mindfulness Interv. acute pain Exp. An Exam. Individ. Differ.*, vol. 1, pp. 1689–1699, 2015.
3. H. Liu and R. Setiono, “A probabilistic approach to feature selection - a filter solution,” *Proc 13th Int. Conf. Mach. Learn.*, vol. 96, pp. 319–327, 1996.
4. L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” *Int. Conf. Mach. Learn.*, pp. 1–8, 2003.
5. E. R. Hruschka and T. F. Covoes, “Feature Selection for Cluster Analysis: an Approach Based on the Simplified Silhouette Criterion,” *Int. Conf. Comput. Intell. Model. Control Autom. Int. Conf. Intell. Agents, Web Technol. Internet Commer.*, pp. 32–38, 2005.
6. D. Karaboga and C. Ozturk, “Neural networks training by artificial bee colony algorithm on pattern classification,” *Neural Netw. World*, vol. 19, no. 3, pp. 279–292, 2009.
7. R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.
8. B. Xue, M. J. Zhang, and W. N. Browne, “Particle swarm optimization for feature selection in classification: a multi-objective approach,” *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–71, 2013.
9. Song.M.H., LeeJ, Cho.S.P., Lee.K.1 and Yoo.S.K (2005), 'Support vector machine Based Arrhythmia classification using reduced Features', International Journal of control, Automation, and Systems, Vol 3, pp.571-579
10. S.-W. Lin, Z.-J. Lee, S.-C. Chen, and T.-Y. Tseng, “Parameter determination of support vector machine and feature selection using simulated annealing approach,” *Appl. Soft Comput.*, vol. 8, no. 4, pp. 1505–1512, 2008.
11. M. Gütlein, E. Frank, M. Hall, and A. Karwath, “Large-scale attribute selection using wrappers,” in *2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings*, 2009, pp. 332–339.
12. M. Esseghir, “Effective wrapper-filter hybridization through grasp schemata,” *MLR Work. Conf. Proc*, vol. 10, no. i, pp. 45–54, 2010.
13. A. Unler, A. Murat, and R. B. Chinnam, “Mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification,” *Inf. Sci. (Ny).*, vol. 181, no. 20, pp. 4625–4641, 2011.
14. Swati S. and Ashok G. (2013), ‘Feature selectionfor medical diagnosis: Evaluation for cardiovascular diseases', Expert Systems with applications 40, pp.4146-4I53.
15. ThananPravit.,AnanBanharnsakun.,BoonsermKaewkamnerdpong. andTiraneeAchalakul. (2013), 'Reducing bioinfomatics data dimension with ABC-KNN', Neurocomputing 116, pp367-381.
16. J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,” Expert Syst. Appl., vol. 40, no. 1, pp. 96–104, 2013.
17. Nguyen, H.V. and Bai, L., 2010, November. Cosine similarity metric learning for face verification. In Asian conference on computer vision (pp. 709-720). Springer, Berlin, Heidelberg.
18. Dua, D. and KarraTaniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
19. Duan, K.B. and Keerthi, S.S., 2005, June. Which is the best multiclass SVM method? An empirical study. In *International workshop on multiple classifier systems* (pp. 278-285). Springer, Berlin, Heidelberg.