A Hybrid Feature Selection Approach for Data Clustering Based on Ant Colony Optimization

Rajesh Dwivedi[1](✉), Aruna Tiwari[1], Neha Bharill[2], and Milind Ratnaparkhe[3]

[1] Indian Institute of Technology Indore, Indore 453552, India
anubhav.dwivedi8@gmail.com, artiwari@iiti.ac.in
[2] Ecole Centrale School of Engineering, Mahindra University,
Hyderabad 500043, India
neha.bharill@mahindrauniversity.edu.in
[3] ICAR-Indian Institute of Soybean Research, Indore 452001, India

Abstract. Machine learning, data mining, and pattern recognition all require feature selection when working with high-dimensional data. Feature selection helps in improving the prediction accuracy and significantly reduces the computation time. The problem is that many of the feature selection algorithms use a sequential search strategy to choose the most important features. This means that each time you add or remove a feature from the dataset, you get stuck in a local optimum. This paper proposes a hybrid feature selection technique based on ant colony optimization that randomly selects features and quantifies their quality using K-means clustering in terms of silhouette index and laplacian score. The proposed hybrid feature selection technique allows for random selection of features, which facilitates a better exploration of feature space and avoids the problem of being trapped in a local optimal solution, while also generating a global optimal solution. Furthermore experimental investigation shows that the proposed method outperforms the state-of-the-art method.

Keywords: Ant Colony Optimization · Jaccard index · K-means clustering · Laplacian Score · Silhouette Index

1 Introduction

Feature selection is widely used in various data mining and machine learning tasks such as classification, clustering, and regression to improve readability and interpretability. Due to the popularity of feature selection in such areas, so far, researchers have primarily attempted to analyze and explain feature selection tasks in supervised learning area, especially in classification, but in unsupervised learning area [15], especially clustering [8] it has not been explored extensively. A feature selection method generally consists of four main phases: selection, examination, terminating criterion, and validation. The first phase involves

M. Tanveer et al. (Eds.): ICONIP 2022, LNCS 13625, pp. 659–670, 2023.
https://doi.org/10.1007/978-3-031-30111-7_55

selecting a feature subset using a predetermined search strategy, like sequential search, sequential floating search, or complete search. The second phase consists of examining the chosen feature subset by a specific criterion. After getting termination criteria, the third step selects the best performing subset from all possible subsets. The last step involves the validation of the chosen subset using the validation metrics.

The remaining portion of this paper is arranged as follows: Sect. 2 provides a literature review of the existing work. The proposed feature selection method based on Ant Colony Optimization is introduced in Sect. 3. Section 4 presents experimental results on various benchmark datasets. Finally, Sect. 5 presents the conclusion and future work.

2 Literature Review

This section presents the various techniques for features selection proposed by other researchers. Dash and Liu [2] developed a hybrid feature selection approach that calculates entropy from the similarity of data and uses a measure based on entropy to evaluate the features in the filter stage. The wrapper stage uses scatter separability criteria and k-means clustering to select the relevant feature subset. A drawback of this approach is its high computation cost. Later on, Hruschka et al. [7] proposed a hybrid feature selection approach that uses a Bayesian filter with k-means clustering to identify the relevant feature subset. They used a Bayesian network that uses Markov blanket property for the filter approach. A drawback of this approach is that they have only tested it for datasets having less than 30 features. By adopting the same idea proposed by Dash and Liu [2], Li et al. [9] proposed a new hybrid feature selection approach, in which they used Fuzzy Feature Evaluation Index (FFEI) with an exponential entropy index to evaluate the feature in the filter stage to increase the performance. They used a scatter separability criterion and Fuzzy C-Means algorithm for the wrapper stage. This approach also suffers from high computation costs.

In 2015 another feature selection approach suggested by Nahato et al. [10] uses rough set theory to identify the relevant features. They used rough indiscernibility relation to select the reducts and trained backpropagation neural networks using the selected reducts. This method was tested on statlog heart disease datasets, wisconsin breast cancer dataset, and hepatitis dataset taken from UCI machine learning library [1] and achieved an accuracy of 90.4%, 98.6%, and 97.3% with 6, 7, and 13 features, respectively. Later on, in 2016 Solorio et al. [14] proposed a hybrid feature selection technique that uses a laplacian score to rank the features and a modified Calinski Harabase index to measure a feature subset. They tested their approach on various benchmark datasets taken from the UCI Machine learning repository [1] and also on several synthetic datasets and thus achieve better results than approaches proposed by Dash and Liu [2] and Li et al. [9].

In 1990, Dorigo et al. [3] came up with the idea of an Ant Colony Optimization (ACO). This approach mimics the social behaviour of ants searching for food.

Initially, it was developed to solve the famous traveling salesman problem. Later it was applied to various complex optimization problems like feature selection [11]. An unsupervised ACO based feature selection was proposed by Tabakhi et al. [17] that uses Cosine similarity measures to measure the similarity between features. In this work, the number of artificial ants used was equal to the number of attributes in the dataset such that each ant was responsible for constructing a feature subset. The frequency of selected attributes on different subsets was used to update the pheromone value. Feature having low similarity, and high pheromone value was added to feature subset in every step till max iteration. They tested their approach on several UCI machine learning datasets [1] like wine and breast cancer datasets and got an average classification error of 19.8%. In 2017, Dhalia et al. [16] proposed another ACO based approach that uses a tandem run strategy to select the relevant feature subset. They used Cosine similarity measure to measure the similarity between features and support vector machine (SVM) for assigning the fitness to a feature and further SVM is used for classification. They tested their approach on Lung CT scan images to diagnose bronchitis and achieved 81.66% accuracy.

From the presented work, it can be inferred that hybrid methods are performing well in comparison of filter and wrapper methods. Also, ACO is used in various tasks for the feature selection and gives an increased accuracy; therefore, in the proposed work, a hybrid feature selection approach based on ACO is presented which uses silhouette index and laplacian score as a fitness measure and gives an increased clustering performance on various benchmark datasets.

3 Proposed Method

In this work, we proposed a novel hybrid feature selection technique based on ant colony optimization [4] (NHFS based ACO) that follows the tandem run strategy [5] to select the best feature subset.

The simulation model is expressed by a completely connected undirected graph G = (V, E) having a one-to-one mapping between vertices and features. Hence the number of vertices (v_n) equals the number of features (f_n). V denotes the set of vertices as v_1, v_2, $v_3 .... v_n$ and E denotes the set of edges (e_1, e_2, $e_3 .... e_n$) joining any two vertices in the graph. In this model, the number of artificial ants (N_{ant}) is taken same as the number of features (f_n) to avoid being trapped in the local optimum, so $f_n = v_n = N_{ant}$. In ACO each artificial ant constructs a feature subset (F_i). The N denotes the set of all feature subsets created by ants and n_{max} indicates the maximum number of features possible in each subset, then $N_{ant} = N$ and $0 \leq n_{max} \leq n$.

To make the feature subsets, each ant starts from a vertex and creates a feature subset by traversing different vertices in between. Every feature is associated with a pheromone value (α) set to a constant initially. n_{tan} denotes the number of features selected by the tandem run strategy.

In this method, three steps are used to choose n_{max} features. In the first step, n feature subsets are made by picking n_{max} features at random. Then, we apply K-means clustering to these subsets and evaluate their efficacy in terms of silhouette index (SI) values. The leader subset is the one with the highest SI value ($g_{bestset}$). Algorithm 1 describes the working of first step. In the second step, n feature subsets are made in a different way, and the selection of n_{max} features is accomplished in three stages. Certain features are chosen randomly (n_{random}), while others ($n_{arbitary}$) are chosen based on their high pheromone and low laplacian scores. On the other hand, some features (n_{tan}) from the leader subset are chosen because they have a high pheromone score but a low laplacian score. Once again, these subsets are used in K-means clustering to evaluate how effective they are in terms of a SI value that is determined in the third step. The subset with the highest SI is called localbest ($l_{bestset}$), and if it is greater than globalbest, it becomes globalbest ($g_{bestset}$). Iterate the second and third steps until max_{iter}. After all iterations, the global bestset is the subset with the highest SI value. The working of second and third step are shown in Algorithm 2.

Algorithm 1

```
Input: Dataset, $n_{max}$
Output: Leader subset after first iteration ($g_{bestset}$)
1: Create $n$ feature subsets, each feature subset will have $n_{max}$ features choosen randomly.
2: These subsets are applied to K-means clustering and the efficacy of these subsets is evaluated in terms of SI value. Take the number of clusters equal to the number of class label as given in the dataset, if class label is not available then decide $k$ randomly.
3: Subset gives the best SI value which is considered as the leader subset and known as $g_{bestset}$.
4: Return $g_{bestset}$.
```

3.1 Computation of Laplacian Score

Laplacian score [6] represents the local preserving power of a feature. It is used for feature ranking in many feature selection approaches. A good feature always has a low laplacian score value. For a given dataset with m instances, a similarity graph is constructed in the form of the weight matrix W of size $m * m$ such that $W = \{w_{11}, w_{12}, w_{13}, ...w_{ij}, ...w_{mm}\}$, where each edge connecting instances x_i to x_j represents similarity in form of weight w_{ij}. Laplacian matrix L is calculated as defined in Eq. (1).

$$L = D - W \tag{1}$$

where D is the diagonal matrix and W is the weight matrix.

Algorithm 2

```
Input: Dataset, $n_{max}$, $max_{iter}$, $g_{bestset}$ after first iteration.
Output: Leader subset ($g_{bestset}$)
Select $n_{max}$ features in $n$ subsets using step 2.
    $n_{random} = n_{max} - n_{remain}$
    $n_{remain} = n_{arbitary} + n_{tan}$.
Select $n_{random}$ features randomly and $n_{arbitary}$ number of features with high pheromone and max heuristic value.
Select $n_{tan}$ features from leader subset with high pheromone and max heuristic value.
These subsets are applied to K-means clustering and the efficacy of these subsets is evaluated in terms of SI value. Take the number of clusters equal to the number of class label as given in the dataset, if class label is not available then decide $k$ randomly.
Subset with maximum SI value is known as localbest.
Compare localbest and globalbest, if localbest if greater than globalbest make localbest as globalbest for further iteration.
Repeat step 1 to 6 till $max_{iter}$.
Return $g_{bestset}$.
```

Let's f_r is the r^{th} feature in all m instances then
$f_r = (f_{r1}, f_{r2}, f_{r3}, f_{r4}, f_{r5}, ...f_{rm})^T$ where $r \in [1, n]$. Laplacian score of f_r is computed as defined in Eq. (2).

$$L_r = \tilde{f_r}^T L \tilde{f_r} / \tilde{f_r}^T D \tilde{f_r} \tag{2}$$

$\tilde{f_r}$ denotes the f_r vector's deviation from the mean and computed as given in Eq. (3).

$$\tilde{f_r} = f_r - (\frac{f_r^T D1}{1^T D1}) \tag{3}$$

where D is the diagonal matrix and $1 = [1, ..., 1]^T$. f_r^T is the transpose of f_r.

After getting laplacian score heuristic value (h_r) is computed using Eq. (4).

$$h_r = \frac{1}{L_r} \tag{4}$$

3.2 Calculation of Pheromone

Step 1: Allocate initial pheromone (α) to all features as given in Eq. (5).

$$\alpha_{f_i} = \frac{1}{n} \tag{5}$$

where $i \in [1, n]$.

Step 2: Each time a feature f_i is selected in a subset F_j, where $i \in [1, n]$ and $j \in [1, n]$ its pheromone update occurs in Eq. (6) and (7).

$$fitness_{f_i} = \frac{SI(F_j)}{n_{max}} \tag{6}$$

$$\alpha_{f_i}(t+1) = \alpha_{f_i}(t) + fitness_{f_i} \tag{7}$$

where $\alpha_{f_i}(t)$ and $\alpha_{f_i}(t+1)$ are pheromones value of a feature f_i at time t and t + 1, respectively.

4 Results and Discussion

4.1 Dataset Details

In the experimental study, we have used three benchmark datasets namely Ionosphere, Sonar and Vehicle silhouettes, which are collected from the UCI Machine Learning repository [1]. Preprocessing of these datasets is performed such as removal of missing values. After preprocessing, details of the datasets are presented in Table 1.

Table 1. Dataset Details

Dataset Name	No. of instances	No. of features	No. of classes
Ionosphere	351	33	2
Sonar	208	60	2
Vehicle silhouettes	813	18	3

4.2 Evaluation Measures

In this approach, two cluster measures and one visualizer are used to evaluate and visualize the clustering performance which are defined as follows:

Jaccard Index (JI). JI [12] is an external evaluation measure for any clustering approach. It evaluates clustering performance based on its similarity to the ground truth or Expert classification. Jaccard index value range from 0 to 1, where 0 represents no match between clustering and ground truth and 1 illustrates a perfect match.

Silhouette Index (SI). SI [13] is known as an internal evaluation measure for any clustering algorithm. It is based on the similarity of a data point within its cluster known as Cohesion and to its nearest cluster known as separation. It ranges from -1 to 1, and a high SI value represents well-clustered data points. The silhouette index is calculated by taking the average of all data point's silhouette coefficients.

Silhouette Visualizer. The silhouette visualizer visualizes which clusters are dense and which are not by displaying the silhouette coefficient for each sample on a per-cluster basis. It also shows how many clusters are achieving the average SI value.

4.3 Parameter Settings for NHFS Based ACO

In this work, experiments are carried out by taking different values of n_{max} for all datasets and for each dataset '50' independent runs of experiments were conducted, therefore $max_{iter} = 50$. Parameters settings for various variables is as follows:

$n_{arbitary} = 30\% of n_{max}$
$n_{tan} = 30\% of n_{max}$
In case of fraction value, consider its ceil value.

4.4 Experimental Analysis

Experiments are performed on the datasets listed in Table 1 and the NHFS based ACO approach is compared with a hybrid feature selection approach developed by Solorio et al. [14] because they used a similar strategy to obtain the best feature subset and measured the results in terms of jaccard index and silhouette index.

Experimental Findings on Ionosphere Dataset: The NHFS based ACO is applied on Ionosphere dataset for different values of n_{max} and results are presented in Table 2.

Table 2. Results on Ionosphere Dataset

Technique used	**No. of feature selected**	**Jaccard Index**	**Silhouette Index**
Solorio et al.	7	0.4376	0.5131
NHFS based ACO	**1**	**0.6132**	**0.7506**
NHFS based ACO	5	0.4486	0.5438
NHFS based ACO	7	0.4589	0.5681

It can be seen from Table 2 that NHFS based ACO is giving increased JI and SI values in comparison to Solorio et al. [14] approach when 1 feature is selected. Both the approaches also worked on 7 number of features, in spite of NHFS based ACO selected more relevant features and that is by performed better than Solorio et al. [14] approach. The silhouette visualizer obtained from Solorio et al. [14] approach shown in Fig. 1(a) showing that some data points having negative silhouette coefficient in blue colored cluster, which shows that those data points are wrongly clustered. On the other hand silhouette visualizer obtained from NHFS based ACO approach shown in Fig. 1(b) shows that all data points have positive silhouette coefficient value and gives better clustering.

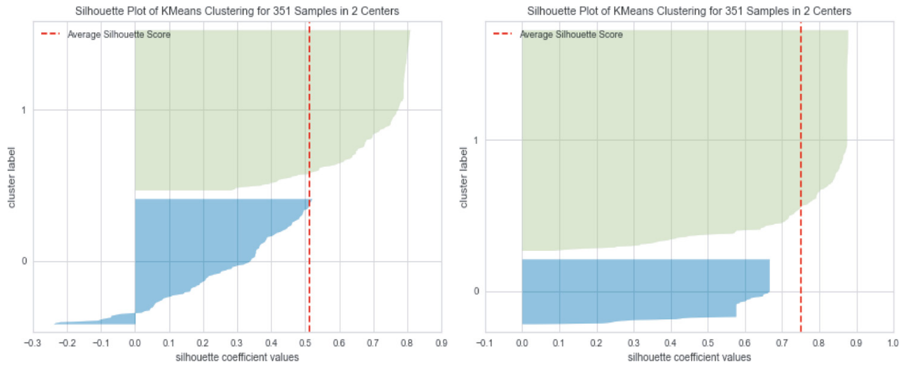

Fig. 1. (a) Silhouette visualizer for Ionosphere using Solorio et al. [14]. (b) Silhouette visualizer for Ionosphere using NHFS based ACO

Experimental Findings on Sonar Dataset: The NHFS based ACO is applied on Sonar dataset for different values of n_{max} and combined results are shown in Table 3.

Table 3. Results on Sonar Dataset

Technique used	**No. of feature selected**	**Jaccard Index**	**Silhouette Index**
Solorio et al.	1	0.3448	0.6304
NHFS based ACO	**1**	**0.4273**	**0.7501**
NHFS based ACO	3	0.4473	0.6319

It can be seen from Table 3 that NHFS based ACO is giving increased JI and SI value in comparison to Solorio et al. [14] approach when 1 feature is getting selected. silhouette visualizer obtained from Solorio et al. [14] and NHFS based ACO are also shown in Fig. 2(a) and Fig. 2(b) to visualize the clustering results.

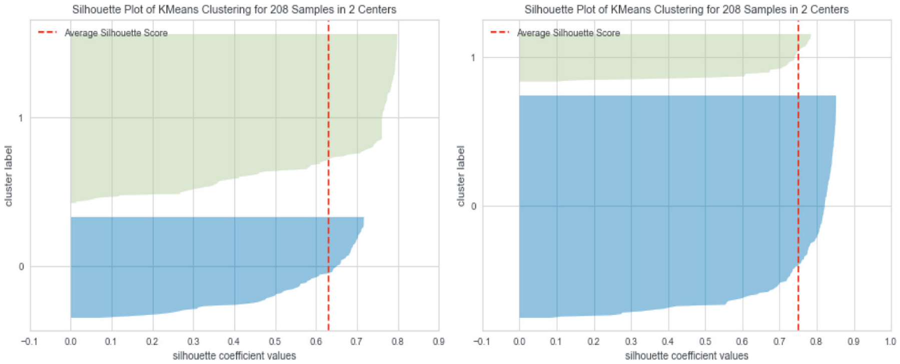

Fig. 2. (a) Silhouette visualizer for Sonar using Solorio et al. [14]. (b) Silhouette visualizer for Sonar using NHFS based ACO

Experimental Findings on Vehicle Silhouettes Dataset: The NHFS based ACO is applied on Vehicle silhouettes dataset for different values of n_{max} and combined results are presented in Table 4.

Table 4. Results on Vehicle silhouettes Dataset

Technique used	**No. of feature selected**	**Jaccard Index**	**Silhouette Index**
Solorio et al.	5	0.2935	0.5635
NHFS based ACO	5	0.3162	0.6603
NHFS based ACO	**4**	**0.3150**	**0.6650**
NHFS based ACO	3	0.3148	0.6562
NHFS based ACO	1	0.3319	0.6516

It can be seen from Table 4 that NHFS based ACO approach is giving increased JI and SI values in comparison to Solorio et al. [14] approach in all cases but the highest SI value when 4 features are taken. Both the approaches also worked on 5 number of features, in spite of NHFS based ACO selected more relevant features and that is by performed better than Solorio et al. [14] approach. Silhouette visualizer obtained from Solorio et al. [14] and NHFS based ACO approach with 4 features are also presented in Fig. 3(a) and Fig. 3(b) to observe the clustering results.

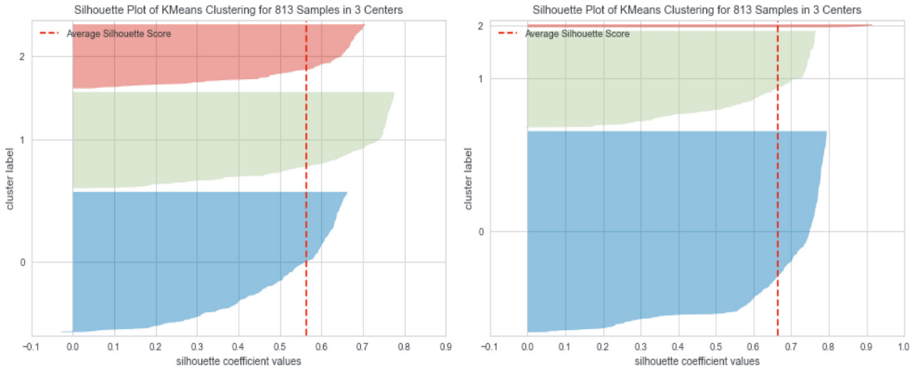

Fig. 3. (a) Silhouette visualizer for Vehicle silhouettes using Solorio et al. [14]. (b) Silhouette visualizer for Vehicle silhouettes using NHFS based ACO

4.5 Comparison

Comparison graphs presented in Fig. 4 showing the comparison between Solorio et al. [14] and NHFS based ACO in terms of silhouette index, shows that NHFS based ACO gives better SI value in all datasets. Whereas Fig. 5 showing the comparison between Solorio et al. [14] and NHFS based ACO in terms of jaccard index shows that NHFS based ACO gives better JI value for all datasets.

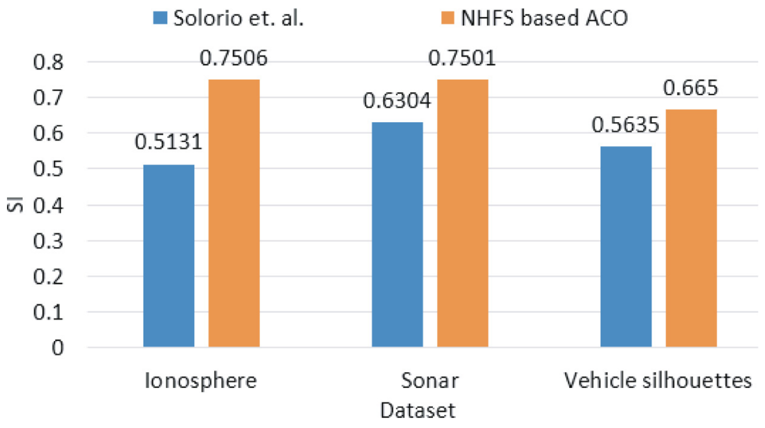

Fig. 4. Comparison between Solorio et al. [14] and NHFS based ACO in SI

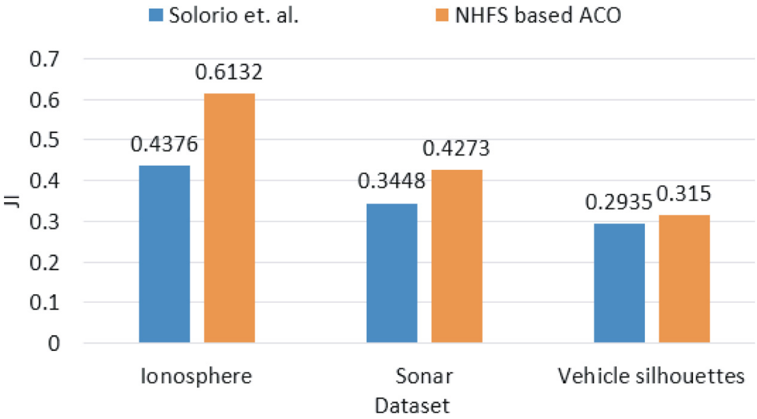

Fig. 5. Comparison between Solorio et al. [14] and NHFS based ACO in JI

5 Conclusion

In the proposed method a hybrid feature selection algorithm based on Ant Colony Optimization named as NHFS based ACO is presented, which removes redundant and irrelevant features that have a negative impact on model building and selects the more appropriate features from data having large number of features. The NHFS based ACO approach uses mixture of laplacian score as well as silhouette index to measure the relevancy of a feature rather than using laplacian score in the filter stages and then silhouette index in the wrapper stage, separately. It also uses tandem run strategy to select the most promising features from the leader subset, which improves the power of proposed approach. The proposed approach is tested on 3 benchmark datasets having a large number of features and achieved the better results than other state-of-the-art approach. The proposed method also worked well on Ionosphere dataset and clustered data points in such a way that all data points have a positive silhouette coefficient. The work on feature selection can be expanded by considering other bio-inspired feature selection algorithms as well by taking various other filter measures.

Acknowledgements. This research is funded by The Council of Scientific and Industrial Research (CSIR), Government of India under grant no. 22(0853)/20/EMR-II.

References

1. Blake, C.: UCI repository of machine learning databases (1998). http://www.ics.uci.edu/~mlearn/MLRepository.html
2. Dash, M., Liu, H.: Feature selection for clustering. In: Terano, T., Liu, H., Chen, A.L.P. (eds.) PAKDD 2000. LNCS (LNAI), vol. 1805, pp. 110–121. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45571-X_13

3. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans. Evol. Comput. **1**(1), 53–66 (1997)
4. Dwivedi, R., Kumar, R., Jangam, E., Kumar, V.: An ant colony optimization based feature selection for data classification. Int. J. Recent Technol. Eng. **7**, 35–40 (2019)
5. Franks, N.R., Richardson, T.: Teaching in tandem-running ants. Nature **439**(7073), 153 (2006)
6. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. Adv. Neural Inf. Process. Syst. **18**, 507–514 (2005)
7. Hruschka, E.R., Covoes, T.F., Ebecken, N.F.: Feature selection for clustering problems: a hybrid algorithm that iterates between k-means and a Bayesian filter. In: Fifth International Conference on Hybrid Intelligent Systems (HIS 2005), pp. 6-pp. IEEE (2005)
8. Kumar, R., Dwivedi, R., Jangam, E.: Hybrid fuzzy C-means using bat optimization and maxi-min distance classifier. In: Singh, M., Gupta, P.K., Tyagi, V., Flusser, J., Ören, T., Kashyap, R. (eds.) ICACDS 2019. CCIS, vol. 1046, pp. 68–79. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-9942-8_7
9. Li, Y., Lu, B.L., Wu, Z.F.: A hybrid method of unsupervised feature selection based on ranking. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 2, pp. 687–690. IEEE (2006)
10. Nahato, K.B., Harichandran, K.N., Arputharaj, K.: Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computat. Math. Methods Med. **2015** (2015). https://doi.org/10.1155/2015/460189
11. Nayar, N., Gautam, S., Singh, P., Mehta, G.: Ant colony optimization: a review of literature and application in feature selection. In: Smys, S., Balas, V.E., Kamel, K.A., Lafata, P. (eds.) Inventive Computation and Information Technologies. LNNS, vol. 173, pp. 285–297. Springer, Singapore (2021). https://doi.org/10.1007/978-981-33-4305-4_22
12. Real, R., Vargas, J.M.: The probabilistic basis of Jaccard's index of similarity. Syst. Biol. **45**(3), 380–385 (1996)
13. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
14. Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A new hybrid filter-wrapper feature selection method for clustering based on ranking. Neurocomputing **214**, 866–880 (2016)
15. Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A review of unsupervised feature selection methods. Artif. Intell. Rev. **53**(2), 907–948 (2020)
16. Sweetlin, J.D., Nehemiah, H.K., Kannan, A.: Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images. Comput. Methods Programs Biomed. **145**, 115–125 (2017)
17. Tabakhi, S., Moradi, P., Akhlaghian, F.: An unsupervised feature selection algorithm based on ant colony optimization. Eng. Appl. Artif. Intell. **32**, 112–123 (2014)