Multi-label Feature Selection using Ant Colony Optimisation with Local Search

Pratham Nayak - 191IT241
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: pratham.191it241@nitk.edu.in

Naveen Shenoy - 191IT134
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: naveenshenoy.191it134@nitk.edu.in

Sarthak Jain - 191IT145
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: sarthak.191it145@nitk.edu.in

***Abstract*—With advancements in technology, it has become easier to collect and store enormous data from which information can be obtained to make predictions on new data. With the emergence of multi-label datasets and multi-label algorithms, it has now become possible to perform multi-label predictions. In multi-label learning, a particular tuple belongs to more than one class label simultaneously. Multi-label learning can be applied across multiple fields. However, this type of learning faces the problem of the curse of dimensionality, that is, the number of samples required in order to estimate an arbitrary function grows exponentially with dimensionality for a given level of accuracy. Therefore the selection of important features is a necessary task. We propose a method of feature selection based on heuristic swarm intelligence. Since feature selection is a combinatorial NP-hard problem, a heuristic algorithm like ant colony optimisation can be used to approximate a good enough solution set. In the proposed method, a solution set is generated such that it has the lowest redundancy and highest relevancy. Standard ACO combined with redundancy and relevancy values is used to generate a solution set. This solution is further improved using random restructure local search that it makes use of multiple classifiers to remove inherent bias. The proposed approach is evaluated on three different multi-label datasets and compared with existing works. The experimental results clearly show that the proposed approach outperforms the existing multi-label feature selection techniques.**

Key words— Ant colony optimisation, feature Selection, local search, multi-label, random restructure.

I. Introduction

Feature selection is a crucial step in the feature engineering process. It involves reducing the number of input variables to produce a predictive model. This is done by eliminating unused or duplicated characteristics between the features. The most crucial features are selected from the list of features. Feature selection is useful for many practical applications, including remote sensing, text categorisation, sequence analysis, image retrieval, etc. The benefits of feature selection are multi-fold. It results in lesser computational power requirements and may prevent the model from over-fitting. The lesser the redundant data, the lesser the influence of noise, thus also improving the accuracy.

The quality of the input features has a significant impact on multi-label categorisation. With the emergence of multi-label datasets and their numerous applications, the domain of features utilised in machine learning and data mining algorithms is expanding quickly. As a result, feature selection methods have become an essential pre-processing step for these algorithms. The feature selection methods can be categorised into integrated, wrapper, or filter-based methods. In contrast to wrapper methods, which use a method where the classification algorithm is used to select the relevant characteristics, filter methods function independently and choose features without consideration of the classification process. The former requires more time but yields more accurate predictions. Since the filter-based methods select the features based on the intrinsic characteristics of the data and do not need to evaluate each generated solution, they are very fast. But this efficiency comes with a loss of effectiveness. Even though the wrapper methods require high computational time, they arrive at an effective solution with better accuracy. Recently, meta-heuristic algorithms have gained wide popularity in solving combinatorial optimisation problems. Examples include the ant colony algorithm, genetic algorithm, particle swarm optimisation and other evolutionary algorithms.

In this work, we propose a feature-selection algorithm for multi-label classification problems using ant colony optimisation along with local search. Wrapper-based feature selection technique is used to further improve the performance over traditional filter-based methods. The following summarises the format of the rest of the paper. In section II, we go through most of the recent work that has been done in the domain of feature selection. After describing the proposed work and model architecture in Section III, we demonstrate the experimental setup and results in Section IV. Finally, we conclude with the inferences obtained in Section V.

TABLE I
SUMMARY OF LITERATURE SURVEY

Authors	Methodology	Merits	Limitations
Paniri et al.	Used ACO for feature selection with feaures as nodes in the graph.	Cosine similarity based initialization along with redundancy and relevance concepts used to get similarity between features and labels.	Used filter based feature selection methods. Evaluated solutions on KNN classifier only.
Jiam et al.	Technique for multi-label feature selection by first lowering the output space's dimension before using feature selection.	Helps to remove the negative effect of the noisy labels in feature selection.	Does not work well when dataset is high dimensional.
Jović et al.	Identified hybrid feature selection methods which are a mixture of evolutionary computation based heuristic algorithms and genetic algorithm.	Found that evaluation of proposed methodologies of feature selection should be made on larger datasets.	Checked feature selection application in only few popular domains.
Kashef and Nezamabadi-pour	First removing the irrelevant features in order to remove number of existing features.	Feature selection process is fast.	Used filter method for algorithm evaluation.

II. LITERATURE SURVEY

An ant colony optimisation-based multi-label feature selection algorithm using redundancy and relevance concepts is proposed in [1]. The authors considered the features as nodes in the ant colony optimisation algorithm. The features with the highest pheromone values are selected as the final set of features. Apart from this, the filter method is used in this work as only after all the iterations are completed the performance of the selected feature set is evaluated using the MLKNN classifier.

Paper [2] used an innovative technique for multi-label feature selection where they first lowered the output space's dimension before using feature selection. The proposed method exploits the latent semantics of the multi-labels. This helps to remove the negative effect of the noisy labels in feature selection. Other than that, it effectively uses label correlations to find shared features. The objective function produced by their final approach is computationally demanding because it must iteratively evolve three matrices at once. When the dataset is high dimensional, this approach does not work well.

This paper [3] focuses on feature selection and gives an overview of the existing approaches that can be used to handle various problem classes. In order to find out which approaches work the best for a given task, it also takes into account the most significant application domains and examines comparative studies on feature selection within those domains. The paper identified hybrid feature selection methods, which are a mixture of evolutionary computation-based heuristic algorithms, for example, the swarm intelligence algorithms like lion search and the genetic algorithm. Also, various application areas were researched, such as image processing and bioinformatics, where features would be present in the form of complex structures. Other than that, approaches such as the filter and wrapper methods were also discussed. This paper also stressed the fact that to get more realistic and reliable results, the proposed methodologies should be evaluated on larger datasets. Apart from this, the work also discussed that feature relevance and redundancy with respect to the objective are the basis for feature set reduction.

A survey on swarm intelligence for feature selection covering some of the common algorithms such as PSO, ABC and ACO is provided in [4]. Along with this, several commonly faced issues and challenges faced were also discussed. Problems such as feature selection bias and computational costs were extensively discussed. [5] suggests using multiple classifiers in the wrapper-based methods rather than using only a single classifier in order to avoid bias due to a particular classifier. The filter-based selection methods generally are less effective than the wrapper-based methods in that they converge to the local optimal solution.

Paper [6] uses the idea of first removing the irrelevant features in order to reduce the number of existing features. Then redundant features are removed using an evolutionary algorithm. Symmetrical uncertainty is used in order to measure the correlation between labels and features. Symmetrical uncertainty is directly proportional to information gain between two variables and inversely proportional to the sum of the entropy of both variables. The method used is a filter based method. Paper [7] uses particle swarm intelligence optimisation using mutual information gain, which is a statistical method used for the measurement of dependency between two variables. The performance of this framework is compared with classification algorithms like SVM, Naive Bayes and Decision Tree.

III. METHODOLOGY

A. Ant Colony Optimisation

Ant colony optimisation (ACO) is a meta-heuristic swarm intelligence algorithm used to solve optimisation problems of combinatorial nature. The algorithm mimics real ants and their foraging and searching behaviour. Chemical substances called pheromones are deposited by the ants on the path they traverse. These pheromones are then used by other ants in their search. The more ants that traverse a path, the stronger the pheromone concentration gets on the path. There is a high probability for an ant to choose a path with a higher pheromone compared to other paths. The pheromone content on the paths constantly changes due to their evaporation as well as due to deposition by other ants. The evaporation of the pheromones prevents the ants from falling into a local optimum. Thus the two concepts, pheromone evaporation and probabilistic selection of paths derived from the real ants' behaviour used in ACO help to solve combinatorial optimisation problems. In addition to the properties possessed by real ants, artificial ants have some additional properties. These include having the ability to store the history of their previous routes and regulating the pheromones on the best-found paths.

B. Ant colony optimisation for feature selection

Feature selection algorithms aim to find an optimal subset from a large set of features. This is done by considering only those features in the final subset that are non-redundant and relevant. ACO is executed on a fully connected graph with the nodes of the graph as the features. Fig. 1 shows how the features are represented in the form of nodes of the graph. Each feature has a corresponding τ (pheromone) value.

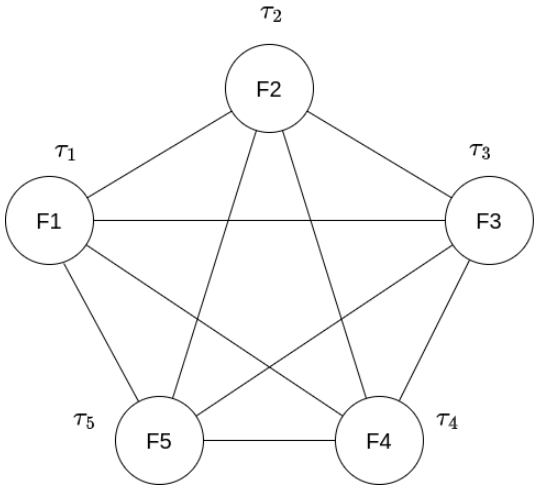

Fig. 1. Graph Representation of Feature Selection using ACO

Initially, an ant is placed on a feature randomly. Based on a pre-decided number of features that are to be included in the final solution, the ant traverses the graph with each new feature on the path being included in the current set of selected features. An iteration consists of several ants completing their traversals. Several such iterations are performed. After each traversal is completed, all pheromones are updated according to the global update rule. Once all iterations are completed, a subset of features is chosen based on decreasing order of pheromone values.

1) Measure feature redundancy and relevance: The redundancy between features is stored in a $d \times d$ matrix $(fCorr)$. The relevancy between features and labels is stored in a $(d \times l)$ matrix $(flCorr)$. The values in the $fCorr$ and $flCorr$ matrices are calculated using the Pearson correlation coefficient between features and label representations. The Pearson correlation coefficient between two n-dimensional feature vectors x and y can be represented using eq. (1).

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2})(\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2})} \tag{1}$$

2) Initialisation of pheromones: For each feature, the pheromone value is initialized as the maximum of the cosine similarity between the feature and different class labels.

$$\cos(A, B) = \frac{AB}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{(\sqrt{\sum_{i=1}^{n} A_i^2})(\sqrt{\sum_{i=1}^{n} B_i^2})} \tag{2}$$

Eq. (2) represents the formula for calculation of cosine similarity and between two vectors. Since the values obtained from above step can be very small, we apply min–max normalization to scale them in the interval [0 , 1] as shown in eq. (3).

$$\tau_{scaled} = \frac{\tau - \tau_{min}}{\tau_{max} - \tau_{min}} \tag{3}$$

3) State Transition Rule: An ant k moves from feature i to feature j according to eq. (4), where q is randomly generated number between 0 and 1.

$$P_{ij}^{k}(t) = \begin{cases} \frac{[\tau_j][\eta_1(F_j)][\eta_2(F_i,F_j)]^\beta}{\sum_{u \in N_i^k}[\tau_u][\eta_1(F_u)][\eta_2(F_i,F_u)]^\beta}, & \forall j \in N_i^k, \text{ if } q > q_0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$j = \underset{u \in N_i^k}{\operatorname{argmax}} \ [\tau_u][\eta_1(F_u)][\eta_2(F_i, F_u)]^\beta, \text{ if } q \leq q_0 \tag{5}$$

Here, N_i^k is the set of all unvisited features when the k^{th} ant is at feature i. $\eta_1(F_j)$ is the relevance factor or the maximum correlation between the feature j and the different class labels $(l \in Y), \eta_1(F_j) = max_l(flCorr_{j,l})$. The factor $\eta_2(F_i, F_j) = 1/fCorr(i, j)$ measures the redundancy and is the inverse of the correlation between feature i and feature j. $P_{ij}^k(t)$ is the probability that the k^{th} ant moves from feature i to feature j at time t. Here, $[\tau_j]$ is the amount of pheromone that is currently present at feature j. Based on the value of the probabilistic variable q, the state transition rule can either lead to an explorative or exploitative search. Explorative search occurs when $q > q_0$ following eq. 4. If $q \leq q_0$, then exploitation occurs, and the next feature is chosen based on

the feature with the maximum value of the function as defined in eq. 5.

4) Global Update: The pheromone global updating rule is applied every time an iteration is completed. The number of times an ant visits a feature is stored in a vector called "visit counter (VC)". At the end of an iteration, say at time t, the pheromones for each feature are updated according to eq. (6).

$$\tau_i(t+1) = (1-\rho)\tau_i(t) + \frac{VC(i)}{\sum_{i=1}^{d} VC(i)} \tag{6}$$

5) Generating and evaluating the Solution: Once all the iterations are completed, generating the solution involves selecting the top features with the highest pheromone content as the set of optimal features. Evaluating the solution using only one classifier makes it vulnerable to inherent bias. Therefore, a variety of classifiers employing different strategies to make predictions are used. The multi-label versions of the following classifiers are used:-

- KNN: It is a non-parametric model that uses majority voting amongst the k-nearest neighbours of a test sample to predict its labels.
- Gaussian Naive Bayes: It is a supervised learning algorithm based on the Bayes theorem assuming conditional independence between the features.
- Logistic Regression: It is used to classify a label using the linear combination of input features.

The KNN model is non-parametric, while Naive Bayes and logistic regression are parametric. KNN supports non-linear solutions, while this is not the case with logistic regression. Gaussian Naive Bayes is a generative model in that it first models the joint distribution of features and labels and then predicts the probability for a data point. On the other hand, logistic regression is a discriminative model in that it directly predicts the posterior probability. Solutions are evaluated using all three of the above classifiers, and the performance is measured as the average of the three accuracies.

6) Improving solution using Local Search: Given a candidate solution, local search attempts to move to a neighbour solution. The neighbour solution is such that it differs from the candidate solution by a minimum extent. We use random restructure local search algorithm. In this method, of the selected features that are part of the generated solution, only half of them are retained. The remaining half of the features are regenerated by applying the state transition rule of the ant colony optimisation. Only exploitative search takes place in this phase, i.e. the next paths are selected according to eq. (5). At each time step, local search is applied to the current best solution. The current best solution is initially the solution generated by the ant colony optimisation. Then the current best solution is updated based on the solution obtained using local search in the previous iteration if and only if it gives better accuracy.

Algorithm 1 Multi-label Feature Selection using Ant Colony Optimisation with Local Search

Require:
D: Dataset with n samples, d features and l labels.
$nIterations$: Number of iterations of ACO
$nAnts$: Number of ants in each iteration
$nFeatures$: Number of features to select
$Evaluate(D, F)$: A function that evaluates the feature subset F of dataset D using Logistic Regression, Naive Bayes and K-Nearest Neighbour and returns the average accuracy.
Output:
F: Set of features of size $nFeatures$
procedure FS(D, $nIterations$, $nAnts$, $nFeatures$)
    $fCorr \leftarrow d \times d$ redundancy matrix as per eq. (1)
    $flCorr \leftarrow d \times l$ relevancy matrix as per eq. (1)
    $\tau \leftarrow$ Initialize as per eq. (2)
    Normalize τ using Min-Max normalization
    for $iterations \leftarrow 1$ to $nIterations$ **do**
        for $ants \leftarrow 1$ to $nAnts$ **do**
            $VC_i \leftarrow 0 \quad 1 \leq i \leq d$
            $Visited \leftarrow Set()$
            $Unvisited \leftarrow Set([1..d])$
            $i \leftarrow$ Random item from $Unvisited$
            $Visited.add(i), Unvisited.erase(i)$
            while $|Visited| < nFeatures$ **do**
                $j \leftarrow$ Assign as per state transition rule
                $Visited.add(j), Unvisited.erase(j)$
            end while
            for $i \in Visited$ **do**
                $VC_i \leftarrow VC_i + 1$
            end for
        end for
        Global update as per eq. (6)
    end for
    Sort features based on decreasing τ values
    $F \leftarrow$ Top $nFeatures$ from the sorted list
    $bestAcc \leftarrow Evaluate(D, F)$
    for $ants \leftarrow 1$ to $nAnts$ **do**
        $Visited \leftarrow$ Randomly sample half of F
        $Unvisited \leftarrow Set([1..d]) - Visited$
        $i \leftarrow$ Random item from $Visited$
        while $|Vistied| < nFeatures$ **do**
            $j \leftarrow$ Assign as per state transition rule
            $Visited.add(j), Unvisited.erase(j)$
        end while
        $F' \leftarrow Visited$
        $acc \leftarrow Evaluate(D, F')$
        if $acc > bestAcc$ **then**
            $bestAcc \leftarrow acc$
            $F \leftarrow F'$
        end if
    end for
    return F

IV. Experimental Results and Analysis

A. Datasets

We used three different datasets for benchmarking the proposed feature selection algorithm. Emotions [8] is a dataset classifying music into 6 emotions, namely amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely and angry-aggressive. A total of 593 instances are present in the dataset. The Image [9] dataset consists of 2000 image instances. In the Image dataset, after moving each image to a separate space, it is divided into 49 blocks using a 7 by 7 grid. The images are then stored in the dataset as 294-dimensional feature vectors. The number of class labels in the Image dataset is 5. Scene [10] is an image dataset containing 2407 images and 6 classes, i.e. beach, sunset, fall foliage, urban, mountain and field. All images are represented as 294-dimensional feature vectors. The characteristics of all datasets have been tabulated in Table II

TABLE II
Characteristics of the benchmark datasets

Dataset	Dataset Size	No. of Features	No. of Labels
Emotions	593	72	6
Image	2000	294	5
Scene	2407	294	6

B. Evaluation Metric

Accuracy is used as the evaluation metric to compare different feature subsets. In the case of multi-label evaluation, 2 different types of accuracies are defined. The first is called subset accuracy, which is defined as the fraction of samples in which all the labels are predicted correctly. The second is where accuracy is defined as the fraction of the correctly predicted labels among all predicted and actual labels. The latter is used in all the performed experiments to evaluate the classifier performance. Mathematically it is expressed using eq. (7)

$$Accuracy(Y, Z) = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{7}$$

C. Parameter Setting

There are a total of 4 parameters that need to be empirically determined. These are ρ, β, $nIterations$ and $nAnts$. Based on the results obtained in [1], $nIterations$ is set to 40, $nAnts$ is set to 25, β is set to 1 and ρ is set to 0.1.

D. Results and Discussion

All the experiments are performed on a Linux-based environment with 13GB Ram and Intel(R) Xeon(R) CPU @ 2.20GHz. The efficiency of the feature selection methods is evaluated by selecting optimal feature sets of different sizes ranging from 10 to 100 or the maximum number of features in the dataset.

MLACO is the feature selection technique proposed in [1]. MLACO-LS1 refers to the modified approach where a single classifier (Multi-label K-nearest neighbours) is used to evaluate the generated solutions during random restructure local search. MLACO-LS3 refers to the modified approach where a combination of 3 different classifiers (Multi-label logistic regression, K-nearest neighbours and Gaussian Naive Bayes) is used to evaluate the generated solutions during the local search.

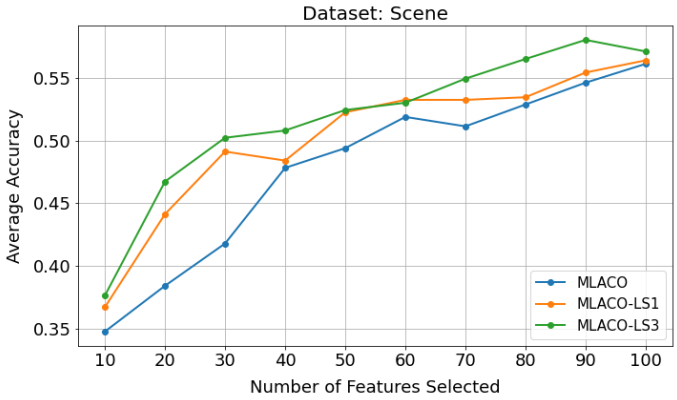

Fig. 2. Average accuracy of classifiers on Scene dataset

Fig. 2 shows the average accuracies obtained for different numbers of selected features when the Scene dataset is used, which has a total of 294 features. Based on the observations, it can be concluded that MLACO gives the least average accuracy. MLACO-LS1 performs better due to the local search optimisation using a KNN-based wrapper. MLACO-LS3 gives the best performance, and this is because a combination of 3 different classifiers removes inherent bias of the classifiers.

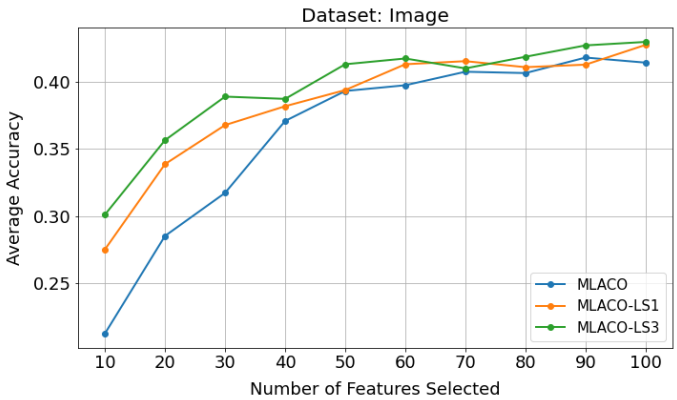

Fig. 3. Average accuracy of classifiers on Image dataset

Fig. 3 shows the average accuracies obtained for different number of selected features when the Image dataset is used, which has a total of 294 features. When the number of selected features is small (< 60), the results are similar to that of the

Scene dataset. However, when a large number of features are selected, the performance of all 3 feature selection techniques is almost identical.

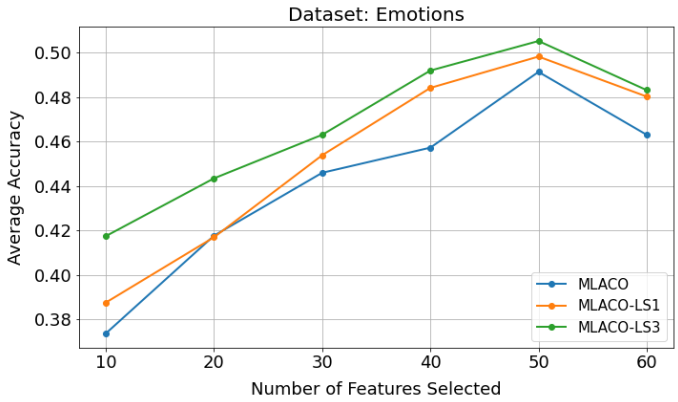

Fig. 4. Average accuracy of classifiers on Emotions dataset

Fig. 4 shows the average accuracies obtained for different number of selected features when the Emotions dataset is used, which has a total of 72 features. As expected, MLACO-LS3 outperforms the other algorithms. However, when the number of selected features is 60, the performance drops. This can be due to increased noise, overfitting or the curse of dimensionality.

TABLE III
AVERAGE ACCURACY OVER 10 FINAL FEATURES' SUBSET

Dataset	MLACO	MLACO-LS1	MLACO-LS3
Scene	0.348	0.367	0.376
Image	0.212	0.275	0.301
Emotions	0.374	0.388	0.417

Table III shows the average accuracy of the 3 classifiers when 10 features are selected from the datasets. As evident, MLACO-LS3 gives the best performance, followed by MLACO-LS1. This shows that local search combined with a multi-classifier based wrapper method helps to significantly improve the quality of the selected features by avoiding local optima through local search and by removing inherent bias by evaluating the generated solution using multiple classifiers of different types.

V. CONCLUSION

The major issue faced in multi-label classification involving datasets with a large number of features is the curse of dimensionality. The number of samples required for the estimation of any arbitrary function grows exponentially with dimensionality therefore, it becomes necessary to select features that are important. Since feature selection is an NP-hard problem, heuristic algorithms can be used to approximate good estimation. The proposed work aims at using an ant colony algorithm-based model. In the proposed approach, each node represents a feature. After applying the standard ACO algorithm, local search optimisation was used to generate new solutions, which were evaluated using multiple classifiers to find the best solution. The method was evaluated over three standard datasets. In order to prevent the algorithm from being biased towards a particular classifier, an average of three different machine learning classifiers was used in evaluating. The proposed method outperforms both standard MLACO as well as MLACO-LS1, suggesting that the performance when using three classifiers is better than using one classifier. ACO combined with multi-classifier-based random restructure local search (MLACO-LS3) gives the best performance.

REFERENCES

[1] Mohsen Paniri, Mohammad Bagher Dowlatshahi, and Hossein Nezamabadi-Pour. "MLACO: A multi-label feature selection algorithm based on ant colony optimization". In: *Knowledge-Based Systems* (2020).

[2] Ling Jiam, Kai Shu, and Huan Liu. "Multi-Label Informed Feature Selection". In: *IJCAI* (2016). ISSN: 1627-1633.

[3] A. Jović, K. Brkić, and N. Bogunović. "A review of feature selection methods with applications". In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2015.

[4] Mengjie Zhang Bach Hoai Nguyen Bing Xue. "A survey on swarm intelligence approaches to feature selection in data mining". In: *Swarm and Evolutionary Computation* 54 (2020), p. 100663. ISSN: 2210-6502.

[5] Mehrdad Rostami et al. "Review of swarm intelligence-based feature selection methods". In: *Engineering Applications of Artificial Intelligence* 100 (2021), p. 104210.

[6] Shima Kashef and Hossein Nezamabadi-pour. "An effective method of multi-label feature selection employing evolutionary algorithms". In: *2017 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*. 2017, pp. 21–25.

[7] Himangshu Shekhar Baruah et al. "A Feature Selection Method using PSO-MI". In: *2020 International Conference on Computational Performance Evaluation (ComPE)*. 2020, pp. 280–284.

[8] Grigorios Tsoumakas et al. "Mulan: A java library for multi-label learning". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2411–2414.

[9] Zhi-Hua Zhou Min-Ling Zhang. "ML-KNN: A lazy learning approach to multi-label learning". In: *Pattern Recognition* 40.7 (2007), pp. 2038–2048. ISSN: 0031-3203.

[10] Matthew R Boutell et al. "Learning multi-label scene classification". In: *Pattern recognition* 37.9 (2004), pp. 1757–1771.