






Original Article

Multi-Label Feature Selection with Graph-based Ant Colony Optimization and Generalized Jaccard Similarity

Sabah Robitan Mahmood ^{a*} , Tahsin Ali Mohammed Amin ^b , Khalid Hassan Ahmed ^a ,
Rebar Dara Mohammed ^b , Pshtiwan Jabar Karim ^c 

^aDepartment of Information Technology, Technical College of Engineering, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.

^bDepartment of Database Technology, Technical College of Engineering, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.

^cDepartment of Computer Science, College of Science, University of Garmian, Kalar, Sulaymaniyah, Iraq.

Submitted: 20 December 2023

Revised: 17 February 2024

Accepted: 1 April 1, 2024

* Corresponding Author:
sabah.robitan@spu.edu.iq

Keywords: Multi-label optimization, Feature selection, Ant Colony, Relevance-redundancy, Generalized Jaccard similarity.

How to cite this paper: S. R. Mahmood, T. A. M. Amin, K. H. Ahmed, R. D. Mohammed, and P. J. Karim, "Multi-Label Feature Selection with Graph-based Ant Colony Optimization and Generalized Jaccard Similarity", *KJAR*, vol. 9, no. 1, pp. 38–51, May 2024, doi: [10.24017/science.2024.1.4](https://doi.org/10.24017/science.2024.1.4).



Copyright: © 2024 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND 4.0)

Abstract: Multi-label learning is a technique that assigns multiple class labels to each data instance. The growth of digital technology resulted in the development of high-dimensional applications in real-world scenarios. Feature selection approaches are extensively used to reduce dimensionality in multi-label learning. The main problems of the recommender system are determining the best match of futures among users but have not engaged with previously. This paper proposes a strategy for selecting features using ant colony optimization (ACO) that incorporates mutual knowledge. The proposed method utilizes ACO to rank features based on their significance. Thus, the search space is mapped to a graph, and each ant traverses the graph, selecting a predetermined number of features. A new information-theoretical metric is introduced to evaluate the features chosen by each ant. Jaccard generalized similarity coefficient is used to select the most suitable communication target for efficient learning outcomes. Mutual information is employed to assess each features relevance to a set of labels and identify redundant features. Pheromones are assigned values based on the effectiveness of the ants in solving the problem. Finally, the features are ranked based on their pheromone values, and the top-ranked features are selected as the final set of attributes. The proposed method is evaluated using real-world datasets. The findings demonstrate that the proposed method outperforms most of existing and advanced approaches. This paper presents a novel feature selection approach for multi-label learning based on ACO. The experimental results confirm the effectiveness of the proposed method compared to existing techniques.

1. Introduction

Data mining is a combination of machine learning statistical analysis, and databases that try to extract meaningful knowledge from data [1]. Recent changes in digital technology may have led to the advancement of big data, which has a high number of dimensions and variables. Selecting features a widely used method to reduce the number of dimensions in data by discarding unnecessary or

duplicate characteristics. For a classification problem, these methods are used to pick a small group of attributes from a large set. In general, feature selection strategies are of two different types: wrappers and filters. Wrapper approaches use a learning model to assess features. When selecting a filter, consider the information-theoretic relevance and duplication of each characteristic. Compared to wrapper approaches, these methods are faster and better for use in the real world because they do not use learning models [1].

Most data mining approaches focus on single-label data, in which just a single label is given to each instance. There are several practical applications, including cancer categorization [2], annotation of images [3], and text categorization [4]. Multiple class labels are provided to each instance. Feature selection techniques are also extensively used for multi-label classification challenges. In contrast with feature selection approaches that just examine a multi-label selection and ignoring the connections between them, assignments require considering the connections that exist within the various labels and features. This makes the selection procedure a more challenging endeavor. Existing ways to choose features can often be put into two groups: issue transformation techniques and adaptive techniques. First, concern transformation techniques to turn the set of labels to only one label. Next, use standard feature selection techniques to pick the features that give the most useful information. Some well-known instances are label power set [5], pruned problem transformation, and binary relevance [6]. With adaptive approaches, the whole space of features and labels is looked through to find a final subset of features. Effective adaptive approaches include the Manifold-based constraint Laplacian score (MCLS) [7], the Maximum reliance and Minimum redundancy (MDMR) [8], and the Multi-variate the Mutual details (MUMI) [9]. The majority of these methods employ greedy search strategies. that cause them to reach a standstill in the local best solution.

In a recent study [10], a rapid method by the name of Multi-label Graph Feature Selection (MGFS) was developed to select important features using multi-label data. This technique begins by using a graph to represent the search space. Next, the PageRank algorithm [11] is applied to data in order to rank the characteristics. Afterward, a subset of top-ranked features is picked. Although it is effective in picking relevant features, this technique overlooks the duplication in the features previously chosen during its search process.

Multi-label learning in high-dimensional information involves assigning many relevant labels to each data point. However, having several features can be harmful, impeding the efficiency of learning algorithms. This paper addresses the issue of convergence problems. Ant Colony Optimization (ACO) algorithms occasionally become trapped and resulting in suggestions that are less than optimal. The pheromone trails may converge on paths that are not optimal for producing varied and pertinent recommendations. It focuses on identifying the most important features efficiently to enhance the performance of multi-label learning algorithms and potentially improve classification accuracy. This study presents supervised multi-label features selection using ACO as a method for selecting multi-label features based on generalized Jaccard similarity. The search space was first represented by an unguided graph using the suggested approach. The features within this graph were denoted by nodes. The weight of each pair of nodes represents the Jaccard similarity score between the characteristics they represent. Next, to rank features based on their relevance to the set of labels, the ACO technique was used, while simultaneously attempting to minimize duplication between characteristics that were similar. Then, each ant's performance is evaluated by a multi-label fitness function that is unique to that ant. The fitness function takes into consideration both the degree to which a particular set of chosen characteristics is pertinent to the label set and the extent to which those features overlap. In addition, no learning model of any kind is used in this function, which is why the approach that has been provided is known as the filter method. Pheromones of features may be kept up to date with the use of this function, which acts as a guide. Therefore, higher pheromones are responsible for assigning those of both important and redundant characteristics. The final subset of features is made up of the characteristics that gained the highest overall scores. Our technique has a number of innovative features, including:

- The method that has been suggested is a multi-label feature selection technique. This method takes into consideration not only how relevant individual characteristics are for a label collection, Additionally, how features that are redundant have been omitted.

- Mutual information is used in the proposed method to evaluate not only how many characteristics are the same, but also how important they are. However, previous approaches, such as [7, 10, 12, 13] simply make use of mutual information (MI) to evaluate the significance of the characteristics. Although this study has selected a method with a high-quality set of features that produce better results than single criterion methods.
- A learning model is employed in the majority of multi-label approaches, when analyzing the features to be selected [14]. Examples of such learning models are machine learning k-nearest neighbor (ML-KNN), machine learning support vector machine, or machine learning naïve bayes. This is done as part of the search phase. Because using any learning model involves a significant amount of computing time, its use in applications that include the current world is restricted. While the wrapper approaches are much slower, the information gain measurement is used in the way that has been presented.
- There have been numerous recent developments in the field of feature selection that make use of the ACO approaches [15-18]. The data sets with a single label were analyzed using each and every one of these techniques.
- To achieve the highest possible level of learning efficacy, the communication object is recommended since it is considered to be the most appropriate based on generalized Jaccard similarity coefficient. The Jaccard similarity metric is used to discover users who have similar tastes by comparing the sets of objects that each user has interacted with. The metric assesses the degree of overlap between the item interactions of users, which is beneficial for collaborative filtering. The generalized Jaccard similarity metric takes into account the count of things that are scored by both users. Its good performance and simplicity have made it commonly utilized in collaborative filtering. However, it does not take into account rating information while estimating similarity. Although Jaccard similarity is widely used, it has several drawbacks. For instance, it does not take into consideration the rating values and tends to yield low similarity values when the number of co-rated items increases.
- Thorough testing on four recognized datasets shows that our strategy produces better results compared to a variety of well-known and cutting-edge multi-label feature selection approaches.

2. Related Works

The goal of multi-label feature selection techniques is to choose informative features that are most relevant to a given set of target labels and have the least amount of overlap with other chosen features. Data transformation and adaptive algorithms are both well-known methods in this area. The objective of the data transformation techniques is to change the feature space with several labels into one with just a single label. For example, in a study a technique known as Binary Relevancy based on Information Gain (BR-IG) was presented [6]. This technique involved translating the binary numbers from the multi-label feature set.

In another study [19], the researchers presented a technique for transformation of data that they referred to as (PPT-MI). Each of these methods does not consider how similar the features and labels are to each other. As a result, these methods are not very precise at choosing features for datasets with more than one label. In contrast to data transformation processes, adaptive approaches were implemented for multi-label data, and as a result, have not experienced any information loss. For example, the MDMR [8] approach first figures out how important. Then, the target label is associated with each feature. It figures out how similar certain features are to each other. The researchers of the paper [9] presented a multivariate filter model that they referred to as MUMI. This model naturally employs shared expertise to calculate the relevance from features among various label sets. The MGFS technique starts by making a feature graph, and then it employs the well-known page ranking method to figure out how important features are based on how they relate to the target labels [10]. However, the assessment technique used by this approach disregards the duplication that exists between the attributes. The authors of [20, 21] suggested a technique for selecting multi-label features by approximately modeling the interdependence between features via the use of mutual information. These methods used mutual information combined with learning models to figure out which features were important and which

ones were the same. Since it relied on a learning model to perform the heavy lifting, this approach is not practical for large-scale data sets like those seen in real life. According to MCLS, a manifold-based scoring system may be used to transform feature space into Euclid label space by utilizing a manifold learning method. Instead of using a step-by-step approach, adaptive methods search through all the available features and labels at once to find the most effective set of features. MCLS is an example of this approach [7].

Multi-label feature selection involves selecting a group of informative features from high-dimensional multi-label data, which is essential for pattern recognition. Conventional multi-label feature selection methods based on information theory use low-order mutual information to assess high-order feature relevance between features and labels [22]. This paper also presents a novel approach named Graph-based Feature Selection Method (GFML) using ACO for choosing features in multi-label scenarios. GFML operates through multiple phases. Initially, it constructs a network, wherein each feature corresponds to a node, and the links between them signify their similarity, which is evaluated using mutual information [22].

In contrast, this study’s proposed approach utilizes mutual information theory and generalized Jaccard similarity to evaluate feature similarity and calculates the relevance of each feature considering the entire label set. Furthermore, it employed ACO to search the solution space and rank attributes based on their importance to the label set while minimizing inter-attribute similarity. For instance, a study [12] introduced Manifold Discriminative Feature Selection (MDFS), which utilizes manifold regularization to leverage label correlations. Recently, a method called MGFS [10] was described as a rapid technique for converting information into a graph and then using the PageRank approach [11] to rank and identify the most important features. In addition, a novel technique for selecting features in multi-label data, known as Multi-Label feature selection algorithm based on ACO (MLACO), is introduced [14]. While this method takes into account both relevancy and redundancy requirements, it lacks a robust metric for picking features and is not suitable for multilabel data in high-dimensional datasets. This study’s proposed solution adopts a filter-based multivariate approach, resulting in significantly improved efficiency compared to wrapper-based techniques [14]. Several previously described approaches for selecting multi-label feature attributes are included in table 1.

Table 1: Strategies for multi-label feature selection, including data transformation, adaptive methods, information gain, mutual information, page rank, and various filters and wrappers.

Method	Year	Evaluation measure	Approach type	Relevancy-Redundancy	Filter/Wrapper
BR-IG [6]	2004	IG	DT	U	F
PPT-MI [19]	2011	MI	DT	U	F
MDMR [8]	2015	MI	AD	M	F
MUMI [9]	2013	MI	AD	U	F
MGFS [10]	2020	PR	AD	U	F
GMFS [20]	2017	MI	AD	M	W
MCLS [7]	2018	Manifold	AD	U	F
GFML [22]	2020	MI+ACO	AD	M	F
Proposed method (JACO)	-	MI+ACO+Jaccard	AD	M	F

3. Materials and Methods

This section describes the specifics of the strategy that is being suggested. The approach that has been suggested is known as ant colony optimization-guided multi-label feature selection using generalized Jaccard similarity ACO (JACO), and it is comprised of three primary components. The first step’s objective is to convert the space of possible solutions into an undirected graph. The nodes in this graph represent the features, and the edges show how the characteristics are linked to each other. In the

second step, the ACO's search procedure is used to select a set of characteristics. The process involves selecting features that are highly relevant but have low redundancy. In the final phase, the features are sorted based on their corresponding pheromone quantities. The highest-scoring features are then chosen for the final feature set. Figure 1 depicts the flowchart of the suggested operational procedure. Additional information on each stage is provided in the section that corresponds to it.

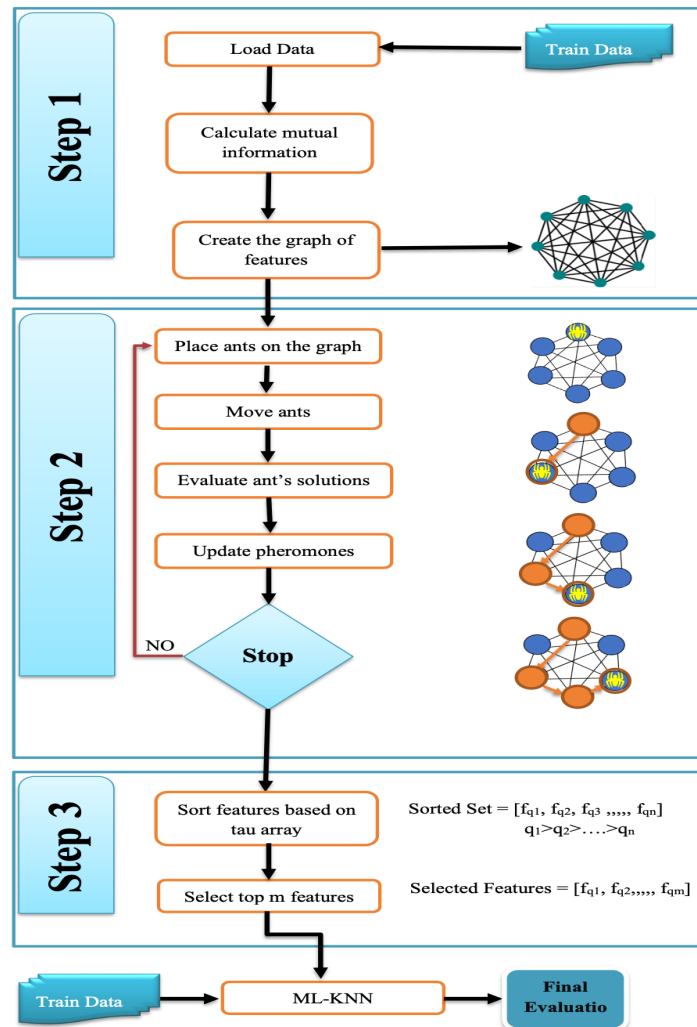


Figure 1: The proposed method's main steps.

3.1. Graph Construction

This section attempts to transform converting the feature space into a graph using the notation. $G = (F, E)$. Every node represents a feature and E represents a group of edges connecting the features. MI metric is used to figure out how similar the nodes are [23]. As illustrated in the following:

$$\frac{sim(A,B) = \sum(a \in A) \sum(b \in B) p(a,b) \log p(a,b)}{p(a)p(b)} \quad (1)$$

Where $p(a)$ and $p(b)$ represent the functions of insignificant distributions of probability of A and B , and $p(a, b)$ demonstrates the combined function of probability of A and B . When there is a lot of information that goes both ways between two variables, it shows that the variables are very dependent on each other. In addition, standardize the data so that it can only have values between 0 and 1. Therefore, any two characteristics that are entirely identical both have similarity values of 1, while any two features that are entirely dissimilar have similarity values of 0. An example of the graph of

characteristics may be seen in figure 2. This graph is undirected, and the weight that is given to each feature in relation to another in this graph is based on the mutual information that is shared between those features. It is also important to note that these nodes are the features, and the only thing that matters is the weight between them; where the features are distributed in the graph is of no significance.

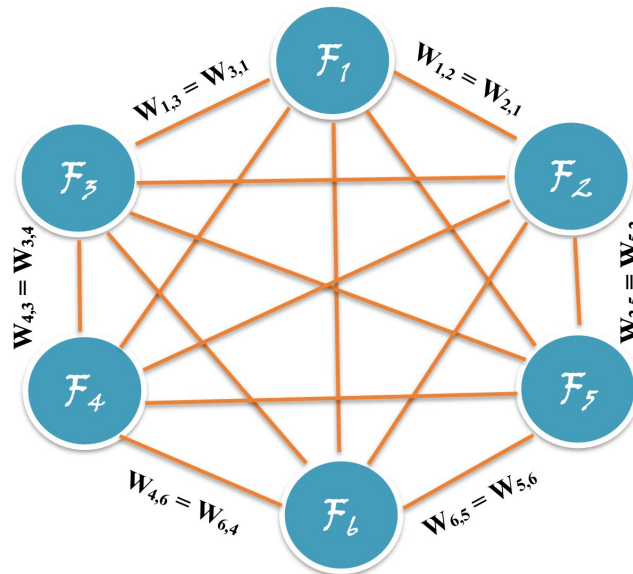


Figure 2: Graphical representation featured for the ACO algorithm.

3.2. Ant Colony Optimization

The ACO [24] draws its motivation from the cooperative nature of ant societies. This technique is based on the collaborative effort of a group of ants working together. Pheromone is the name of the chemical substance that the agents leave behind when they want to connect with one another. The heuristic information in this case is defined as the reverse of the similarity that exists among each pair of the characteristics. In addition, established a desirability metric $\forall i=1 \dots n$, that named "pheromone". This value is allotted to each feature, and the ants are responsible for keeping it up to date. Before any initialization can happen, the initial values of a number of variables, such as ρ (the parameter of pheromone evaporation), m (the final set of attributes), with others, have to be set. These characteristics have the least similarity to be previously chosen features (duplication), the greatest dependence on a set of labels (relevance), and the greatest pheromone values. The "antscore" metric (Eq. (3)) is defined for each round when the population of the ants have picked the number of characteristics that were determined. Using the "Relevancy Redundancy Rule" (RRR) for each feature identified by the ant, the "antscore" value is determined for each ant. The following formula is used to get the RRR for each iteration's features:

$$RRR(Fa) = \sum(li \in L) \sum(fi \in S \text{ } fa \neq fi) \alpha MI(Fa, li) - \gamma MI(Fa, Fj) \quad (2)$$

Where L represents the compilation of all labels, and li is the i th label in L . S represents the compilation of attributes that were selected by each individual ant. Fj is the j th characteristic. $MI(Fa, li)$ is the mutual information among both the a th feature and the i th label. $MI(Fa, lj)$ is the mutual information among a th feature and j th feature. These are two factors that help figure out how important relevance is compared to redundancy. Therefore, the "antscore" is determined as follows:

$$antscore(k) = \sum(fi \in S) RRR(Fi) \quad (3)$$

$\Delta\tau$ is used to establish a measure for each feature. When it is combined with the " $antscore$ " of the ants that have chosen the feature, it yields the final score for that feature.

$$\Delta\tau(F_i) = \sum(k \in C)antscore(k) \tag{4}$$

Where " F_i " refer to the picked feature, an " C " refers to the group of ants that were chosen " F_i ." In the preceding, after the iterations completed, the "pheromone update rule" (Eq. (7)) was applied to update the amount of pheromone visible in each node. The value in this formula represents the entire quantity of pheromone absorbed by each feature as a result of the pheromones left behind by all of the ants that visit the feature. It is important to take notice of the fact that ants have a propensity to offer more pheromones to nodes with higher values for features. Following that, the features were ranked based on the pheromone values they contain, and the top m characteristics were chosen to be included in the final feature set. In addition, every ant determined which step to be taken by solving the problem using either equation 5 or equation 6. It is defined as a greedy technique for selecting the next feature when the k th ant was positioned on the i th feature:

$$j = \arg \max\{\tau_v[\vartheta[(F_i, F_j)]]^\beta\}, \text{ if } q \leq q_0 \text{ } v \in U_i^k \tag{5}$$

Where U_i^k is the quantity of unvisited features, τ_v is the pheromone of the v -th feature, $\vartheta(F_i, F_j) = 1/MI(F_i, F_j)$ is the similarity between two features and determines the importance of pheromone versus similarity ($\beta > 0$), q is a random value between, and q_0 is a constant criterion that is the difference.

In addition to this, equation (6) illustrates a probabilistic method for selecting the subsequent attributes. This formula helps to figure out a probability value for each unvisited characteristic j in the following way:

$$P(i, j) = \{(\tau_j[\vartheta[(F_i, F_j)]]^\beta) / (\sum(v \in J_i^k) \tau_v[\vartheta[(F_i, F_j)]]^\beta), \text{ if } q \leq q_0, \text{ otherwise } 0 \tag{6}$$

The formulas 5 and 6 were determined by the values of the parameters q and q_0 . In the event that q is greater than q_0 , the greedy equation is picked; else, equation 6 is used. The following is how the phrase "*pheromone updaterule*" is defined:

$$\tau_i(t + 1) = (1 - \rho)\tau_i(t) + \Delta\tau(i) \tag{7}$$

Where ρ regulates the rate at which the pheromone evaporates.

$$EJ(X, Y) = \frac{A.B}{\|A\|^2 + \|B\|^2 - A.B} \tag{8}$$

When comparing the contrasts and similarities among finite sample sets, the generalized Jaccard similarity coefficient is a useful tool to utilize. If the value of the Jaccard coefficient is high, then the samples are quite similar to one another. An expansion of the Jaccard coefficient, which is also referred to as the generalized Jaccard, was developed. The way it manifests itself is (8). N -dimensional vectors A and B are used in this equation. It's a vector product when $A.B$. The difficulty with the approach that has been suggested is $O(pn^2 + |iterations| \cdot |ants| \cdot |F|)$.

$$\|A\|^2 \text{ is the norm of vectors } \|A\|^2 = \sqrt{\sum_{i=1}^n a_i^2} \quad (9)$$

4. Results

The proposed method is evaluated on five well-known real-world datasets and compared to a set of methods that include MGFS [10], MDFS [12], MLACO [14], MCLS [7], and GFML [22]. It uses ML-KNN [13] with $k = 40$ to compare the classification performance of the results in terms of ranking loss (RL), hamming loss (HL), average precision (AP), and one error (OE) evaluation measures [25]. In addition to this, the outcomes are an average obtained from 10 separate iterations, with each dataset having a distinct number of characteristics used for each method. Moreover, four multilabel assessment metrics, including hamming loss, ranking loss, average accuracy, and coverage [25] were employed to demonstrate the efficacy of the approaches. Python, a general-purpose programming language, was used to implement each technique on an Apple M1 central processing unit with 8 gigabytes of random-access memory.

4.1. Datasets

In the process of the trials, four datasets were retrieved from the Mulan1 repositories to conduct an analysis of the suggested technique. Table 2 contains a summary of the information pertaining to these dataset’s individual components.

Table 2: Multi-label datasets were used in the assessments.

Name	Patterns	Feature	Label	Label Density	Cardinality Labels
Arts	5000	462	26	0.063	1.636
Business	5000	438	30	0.074	1.470
Education	5000	550	33	0.044	1.461
Health	5000	612	32	0.052	1.662
Science	5000	743	40	0.036	1.451

The maximum number of repetitions that could occur in the experiments was established as 20, and the starting values of the pheromones were each given a value of 0.1 ($\tau_i = 0.1$). Other configurable parameters:

- ρ (rho): 0.1 (presumably the evaporation rate of the pheromones)
- α (alpha): 0.6 (controls the influence of pheromone trails on ant movement)
- β (beta): 1 (controls the influence of heuristic information on ant movement).

4.2. Experiments

This study’s method (JACO) can look at both redundancy of efforts and relevance at the same time, whereas other methods rarely look at either one or both. Tables 3 to 6 demonstrate clearly how JACO compares to alternative modeling techniques on the specified datasets for each evaluation metric. The highest deals are in bold. The values of hamming loss of the approaches that rely on the ML-KNN classifier are shown in table 3. The classification does a much better job overall compared to the other techniques of performance comparison.

Table 3: hamming loss values derived from multi-label feature selection approaches: a comparison.

Dataset	Metrics	features	MGFS	MDFS	MLACO	MCLS	GFML	OURs (JACO)
Arts	HL	40	0.0617	0.063	0.0622	0.0633	0.0621	0.0603
Business	HL	40	0.0288	0.0291	0.029	0.0295	0.0289	0.028
Education	HL	40	0.04142	0.04264	0.04221	0.0448	0.04145	0.04
Health	HL	40	0.0439	0.0437	0.0454	0.0456	0.0444	0.043
Science	HL	40	0.0356	0.0348	0.0352	0.0349	0.034	0.0399

Table 3 displays the results from five separate datasets: Arts, Business, Education, Health, and Science. The table compares the hamming loss values for each dataset using six different methods: MGFS, MDfs, MLACO, MCLS, GFML, and JACO. Based on the results, JACO looks to have the lowest hamming loss value across all datasets, implying that it could be the best performing feature selection approach.

Table 4: Average precision values derived from various multi-label feature selection approaches: a comparison.

Dataset	Metrics	features	MGFS	MDfs	MLACO	MCLS	GFML	OURs (JACO)
Arts	AP	40	0.4726	0.4796	0.4532	0.4408	0.4636	0.4797
Business	AP	40	0.862	0.8673	0.8725	0.8654	0.8714	0.8811
Education	AP	40	0.5293	0.5224	0.5099	0.5002	0.5404	0.5402
Health	AP	40	0.6725	0.6686	0.6627	0.6548	0.6784	0.6797
Science	AP	40	0.4728	0.4795	0.4615	0.4485	0.4745	0.4795

Table 4 compares the AP values obtained from six distinct feature selection approaches across five datasets: Arts, Business, Education, Health, and Science. AP is a metric that assesses the quality of recommendations. JACO has the highest AP score across all five datasets, indicating that it outperforms the other feature selection methods.

Table 5: Analysis of one error values obtained from various multi-label feature selection techniques.

Dataset	Metrics	features	MGFS	MDfs	MLACO	MCLS	GFML	OURs (JACO)
Arts	OE	40	0.6758	0.6712	0.6791	0.6971	0.6818	0.6635
Business	OE	40	0.1298	0.1384	0.1283	0.1367	0.1287	0.1288
Education	OE	40	0.5672	0.596	0.5767	0.605	0.5762	0.5732
Health	OE	40	0.416	0.4077	0.4284	0.4424	0.3979	0.3853
Science	OE	40	0.676	0.6712	0.6988	0.7053	0.6715	0.669

One error is a metric for assessing the performance of a classification model. It denotes the percentage of data points that the model classified inaccurately. In the context of feature selection, a lower one error value suggests that the features chosen by the approach lead to a better classification model. Table 5, compares the one error values produced from six distinct feature selection approaches across five datasets: Arts, Business, Education, Health, and Science. Here's a summary of the results for each dataset, including which feature selection approach had the lowest one error rate:

- **Arts:** JACO (0.6635) outperforms all other methods (MDfs: 0.6712, MLACO: 0.6791, MCLS: 0.6971, GFML: 0.6818, MGFS: 0.6758).
- **Business:** MDfs (0.1287) outperforms all other methods (MGFS: 0.1298, MLACO: 0.1283, MCLS: 0.1367, GFML: 0.1288, JACO (labeled OURs(JACO))): 0.1288).
- **Education:** JACO (0.5732) outperforms all other methods (MGFS: 0.5672, MDfs: 0.5960, MLACO: 0.5767, GFML: 0.5762, MCLS: 0.5732).
- **Health:** JACO (0.3853) outperforms all other methods (MGFS: 0.4160, MDfs: 0.4077, MLACO: 0.4284, MCLS: 0.4424, GFML: 0.3853).
- **Science:** JACO (0.6690) outperforms all other methods (MGFS: 0.6760, MDfs: 0.6712, MLACO: 0.6988, MCLS: 0.7053, GFML: 0.6690).

Table 6: Analysis of the ranking loss results retrieved from various multi-label feature selection approaches.

Dataset	Metrics	features	MGFS	MDfs	MLACO	MCLS	GFML	OURs (JACO)
Arts	RL	40	0.1853	0.1825	0.188	0.198	0.196	0.1827
Business	RL	40	0.04562	0.04817	0.04734	0.04781	0.04559	0.045
Education	RL	40	0.1028	0.1007	0.113	0.1285	0.1138	0.101
Health	RL	40	0.0658	0.0657	0.0649	0.0679	0.0624	0.061
Science	RL	40	0.1395	0.1384	0.1386	0.1399	0.1375	0.1351

Ranking loss is a metric that assesses the quality of a machine learning model's rankings. In the context of feature selection, a lower ranking loss value suggests that the features chosen by the approach contribute to a model capable of producing higher rankings. The table compares the ranking loss values produced from six distinct feature selection approaches across five datasets: Arts, Business, Education, Health, and Science. The results reveal that JACO outperforms all five datasets in terms of feature selection using the ranking loss metric.

The ranking loss criterion and the one error criterion of this study's ML-KNN classifier-based technique are displayed in Tables 4 and 5, correspondingly. The proposed model clearly achieved the lowest value when it is compared to the other techniques across the board. The results of the JACO and the other datasets optimal values are also not that different, and this approach of this study yielded the second-best outcome on these datasets (Table 6).

From figures 3 to 7, the results of the experiments documented the effect of a larger features number on the efficiency of the algorithms. Each figure contains five lines: proposed method values are represented with the black line, while the other four lines were colored to represent the alternative approaches. Figure 3 demonstrates that comparing to the other approaches, the JACO performed better than the other ones

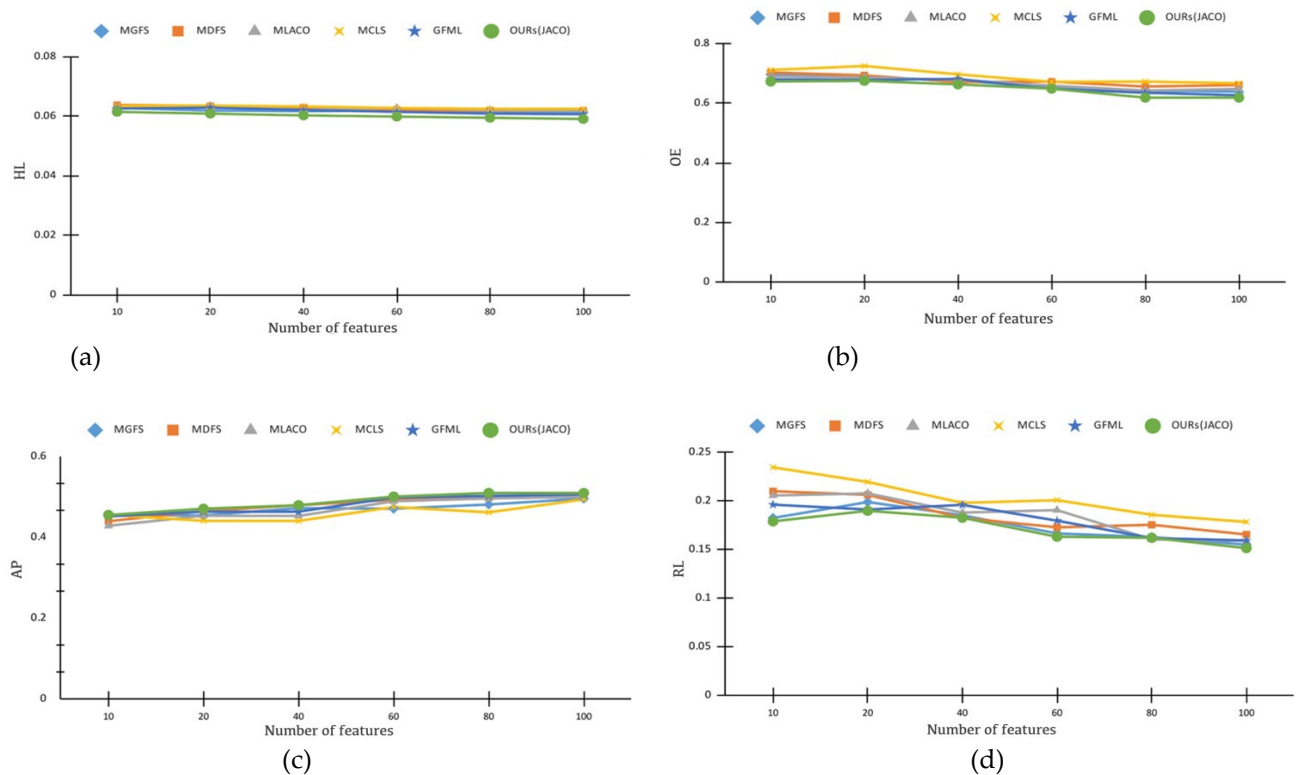


Figure 3: Methodological comparisons on the Arts dataset using (a) HL, (b) OE, (c) AP, and (d) RL metrics.

In figure 4, the JACO outperformed all competing methods with respect to ranking loss, hamming loss, and one error measure in the Business dataset. The remaining metrics on this dataset show that JACO is superior to other methods, while almost matching the performance of MLACO.

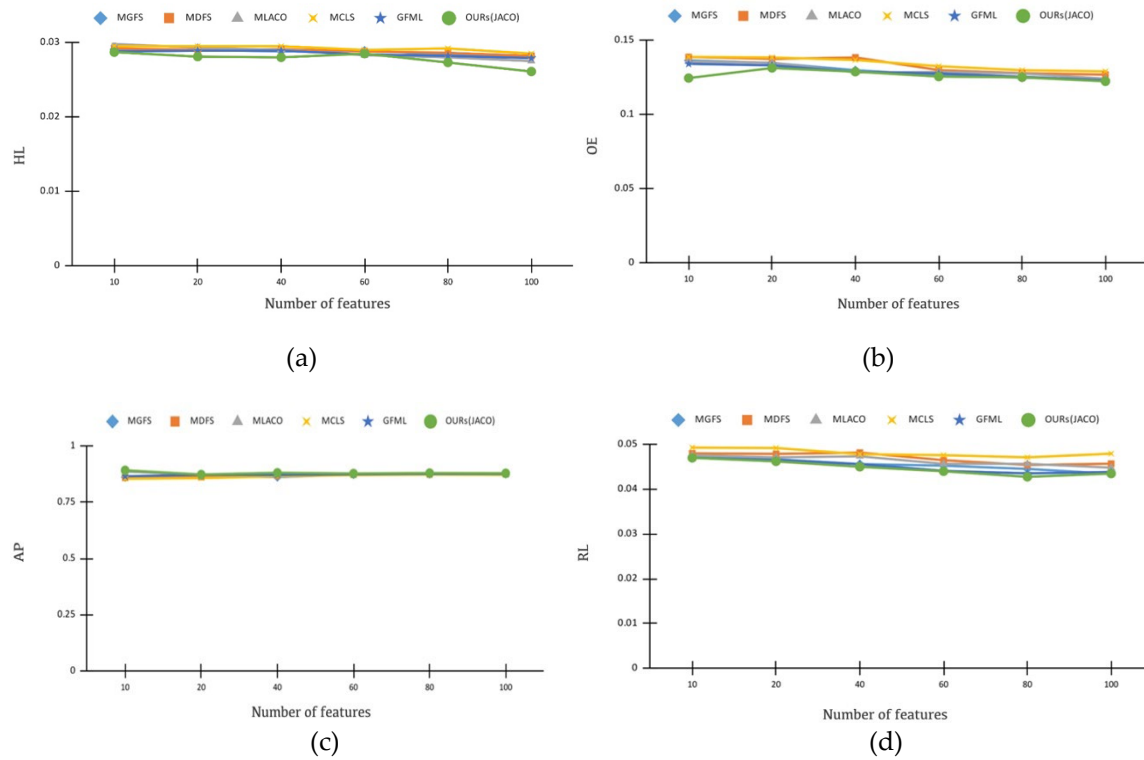


Figure 4: Techniques on the Business dataset are compared with respect to (a) HL, (b) OE, (c) AP, and (d) RL metrics.

Figure 5 illustrates the findings from the Education dataset. The approach of this study outperformed the others in terms of hamming loss and average precision, as shown by the results. In addition, the proposed approach outperformed MGFS on the one error measure.

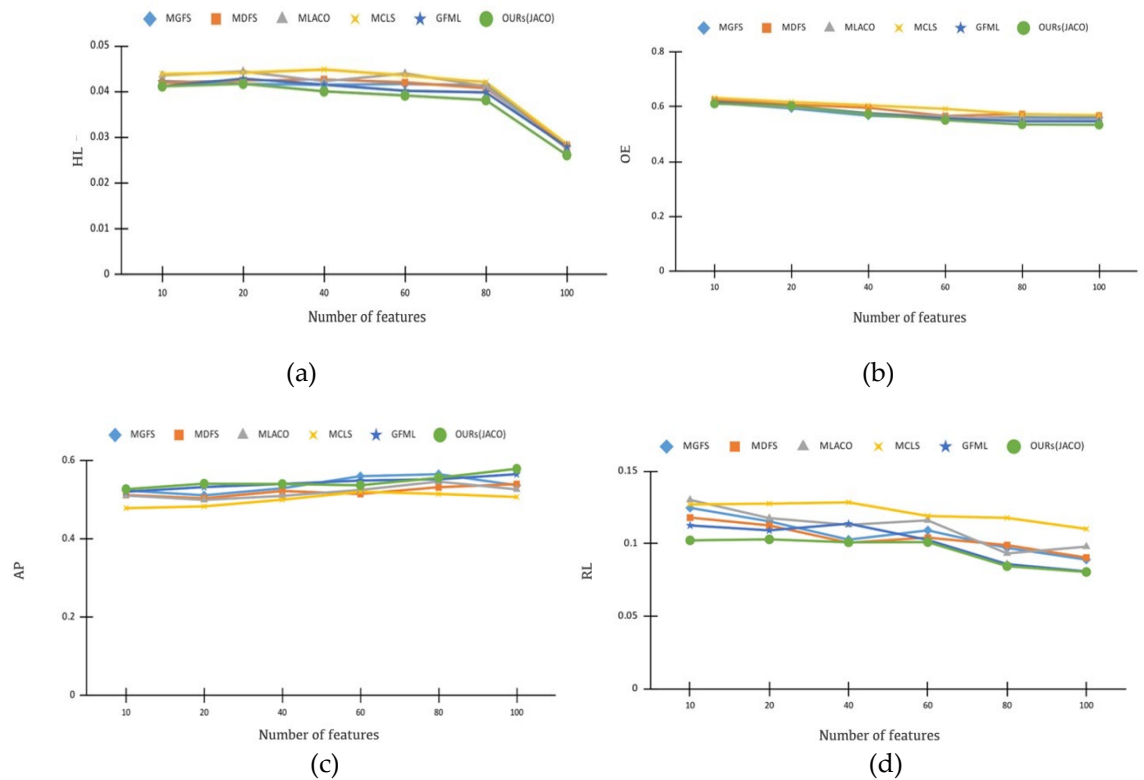


Figure 5: Methods on the education dataset were compared with respect to (a) HL, (b) OE, (c) AP, and (d) RL metrics.

Figure 6 illustrates the findings from the health dataset. The approach of this study outperformed the others in terms of hamming loss and average precision, as shown by the results. In addition, the proposed approach outperforms MGFS on the one error measure.

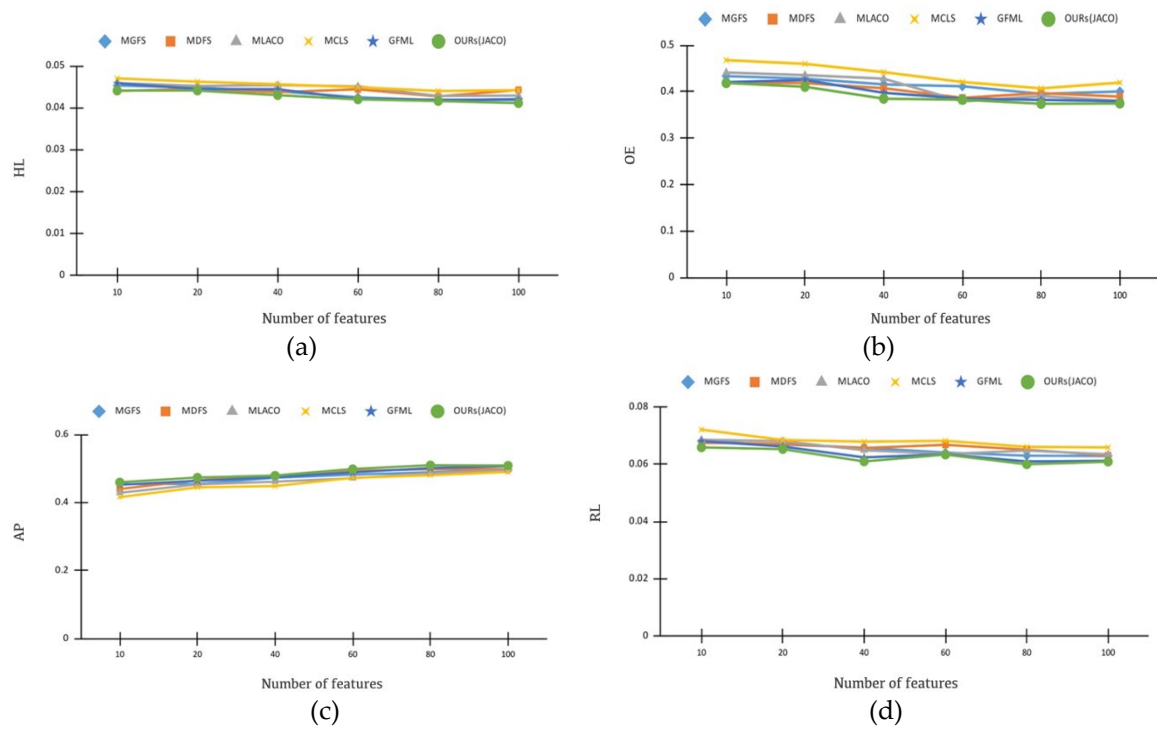


Figure 6: Methods on the Health dataset are compared with respect to (a) HL, (b) OE, (c) AP, and (d) RL metrics.

Figure 7 illustrates the findings from the science dataset. The approach of this study outperformed the others in terms of hamming loss and average precision, as shown by the results. In addition, the proposed approach outperforms MGFS on the one error measure.

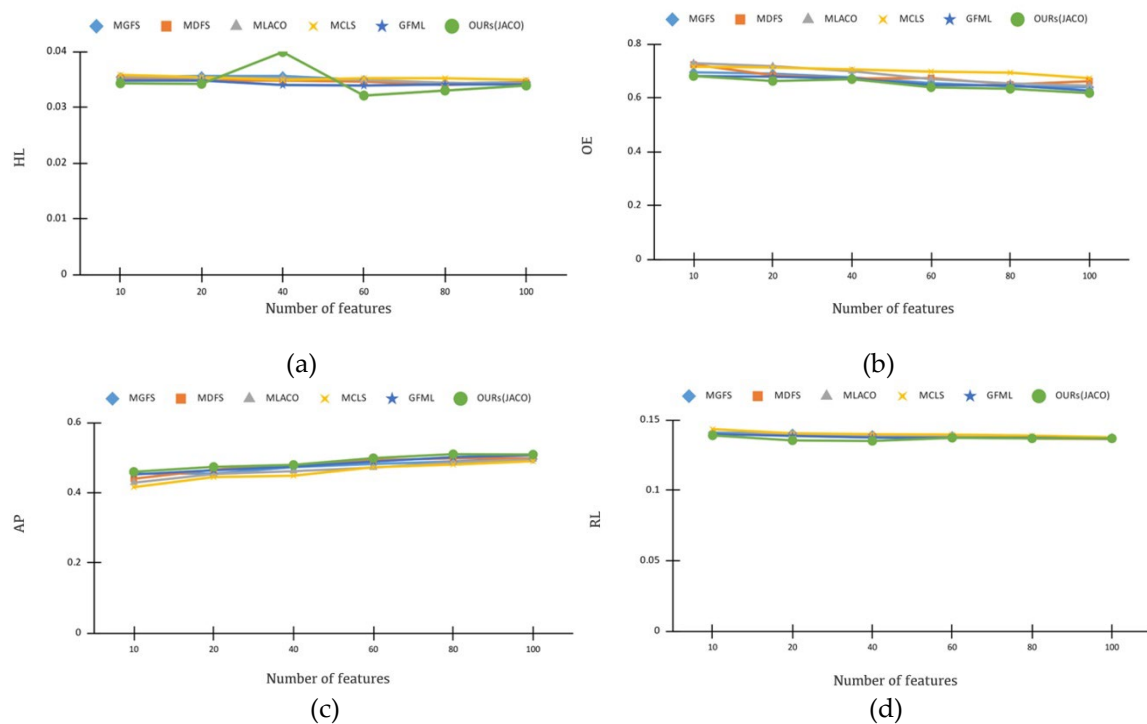


Figure 7: Methods on the Science dataset are compared with respect to (a) HL, (b) OE, (c) AP, and (d) RL metrics.

5. Discussion

It is emphasized that JACO demonstrates comparable or superior performance compared to alternative methods across various datasets and evaluation metrics, as depicted in figures 3 to 7 illustrating the impact of increased feature count on algorithm efficiency. For a better evaluation each metric was tested using different number of features ranging from 10 to 100. On the Business dataset, JACO exhibited superior performance in ranking loss, hamming loss, and one error measure compared to all other methods. As for the Arts dataset, it demonstrated better result across all of other methods with regard to one error, average precision, and hamming loss metrics. However, MDFS showed better ranking loss result. On the Education dataset the proposed method surpassed all of the MGFS, MDFS, MLACO, and MCLS for all metrics, while GFML was better in terms of average precision. Similarly, for health, and science datasets, JACO surpassed competitors in terms of hamming loss and average precision.

In summary, the work suggests that JACO may represent a superior choice for feature selection based on the outcomes of experiments conducted across business, arts, education, health and science datasets.

6. Conclusions

This paper proposes a practical approach for selecting multi-label attribute values. The suggested technique employs using a set of labels, generalized Jaccard similarity and mutual information theory that determines the similarity of features and calculates the relevance of each feature. Furthermore, it uses optimization of ant colonies to rank attributes by searching across the solution space to reduce how similar the attributes are to each other. Rather than relying on a learning model, the suggested solution employs a filter and multivariate approach. As a result, it is much quicker than wrapper-based techniques. To further its classification, it considers both relevance and duplication in its search process as a multivariate technique. There are several ways in which the present project might be expanded in the future. The ACO's search functionality can be enhanced by clustering the graph and putting similar items together. It is also possible to come up with new ways to judge the usefulness and redundancy of certain characteristics in multi-class labels.

Authors contributions: Sabah Robitan Mahmood: Full contribution. Tahsin Ali Mohammed Amin: Full contribution. Khalid Hassan Ahmed: Full contribution. Rebar Dara Mohammed: Full contribution. Pshtiwan Jabar Karim: Full contribution.

Data availability: Data will be available upon reasonable request.

Conflicts of interest: The authors declare that they have no known Competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors did not receive support from any organization for the submitted work.

Reference

- [1] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005, doi: 10.1109/TKDE.2005.66.
- [2] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," *Eng. Appl. Artif. Intell.*, vol. 70, pp. 25–37, Apr. 2018, doi: 10.1016/j.engappai.2017.12.014.
- [3] P. Zhu, Q. Xu, Q. Hu, C. Zhang, and H. Zhao, "Multi-label feature selection with missing labels," *Pattern Recognit.*, vol. 74, pp. 488–502, Feb. 2018, doi: 10.1016/j.patcog.2017.09.036.
- [4] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification," *Appl. Soft Comput.*, vol. 62, pp. 203–215, Jan. 2018, doi: 10.1016/j.asoc.2017.09.038.
- [5] E. A. Cherman, M. C. Monard, and J. Metz, "Multi-label Problem Transformation Methods: a Case Study," *CLEI Electron. J.*, vol. 14, no. 1, Apr. 2011, doi: 10.19153/cleiej.14.1.4.
- [6] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004, doi: 10.1016/j.patcog.2004.03.009.
- [7] R. Huang, W. Jiang, and G. Sun, "Manifold-based constraint Laplacian score for multi-label feature selection," *Pattern Recognit. Lett.*, vol. 112, pp. 346–352, Sep. 2018, doi: 10.1016/j.patrec.2018.08.021.
- [8] Y. Lin, Q. Hu, J. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocomputing*, vol. 168, pp. 92–103, Nov. 2015, doi: 10.1016/j.neucom.2015.06.010.

- [9] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognit. Lett.*, vol. 34, no. 3, pp. 349–357, Feb. 2013, doi: 10.1016/j.patrec.2012.10.005.
- [10] A. Hashemi, M. B. Dowlatshahi, and H. Nezamabadi-pour, "MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality," *Expert Syst. Appl.*, vol. 142, p. 113024, Mar. 2020, doi: 10.1016/j.eswa.2019.113024.
- [11] R. S. Wills, "Google's pagerank: The math behind the search engine," *Math. Intell.*, vol. 28, no. 4, pp. 6–11, Sep. 2006, doi: 10.1007/BF02984696.
- [12] J. Zhang, Z. Luo, C. Li, C. Zhou, and S. Li, "Manifold regularized discriminative feature selection for multi-label learning," *Pattern Recognit.*, vol. 95, pp. 136–150, Nov. 2019, doi: 10.1016/j.patcog.2019.06.003.
- [13] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007, doi: 10.1016/j.patcog.2006.12.019.
- [14] M. Paniri, M. B. Dowlatshahi, and H. Nezamabadi-pour, "MLACO: A multi-label feature selection algorithm based on ant colony optimization," *Knowl.-Based Syst.*, vol. 192, p. 105285, Mar. 2020, doi: 10.1016/j.knosys.2019.105285.
- [15] P. Moradi and M. Rostami, "Integration of graph clustering with ant colony optimization for feature selection," *Knowl.-Based Syst.*, vol. 84, pp. 144–161, 2015, doi: <https://doi.org/10.1016/j.knosys.2015.04.007>.
- [16] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," *Eng. Appl. Artif. Intell.*, vol. 32, pp. 112–123, Jun. 2014, doi: 10.1016/j.engappai.2014.03.007.
- [17] H. Ghimatgar, K. Kazemi, M. S. Helfroush, and A. Aarabi, "An improved feature selection algorithm based on graph clustering and ant colony optimization," *Knowl.-Based Syst.*, vol. 159, pp. 270–285, Nov. 2018, doi: 10.1016/j.knosys.2018.06.025.
- [18] Z. Manbari, F. Akhlaghian Tab, and C. Salavati, "Fast unsupervised feature selection based on the improved binary ant system and mutation strategy," *Neural Comput. Appl.*, vol. 31, no. 9, pp. 4963–4982, Sep. 2019, doi: 10.1007/s00521-018-03991-z.
- [19] G. Doquire and M. Verleysen, "Feature Selection for Multi-label Classification Problems," in *Advances in Computational Intelligence*, vol. 6691, J. Cabestany, I. Rojas, and G. Joya, Eds., in *Lecture Notes in Computer Science*, vol. 6691, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 9–16. doi: 10.1007/978-3-642-21501-8_2.
- [20] F. Li, D. Miao, and W. Pedrycz, "Granular multi-label feature selection based on mutual information," *Pattern Recognit.*, vol. 67, pp. 410–423, Jul. 2017, doi: 10.1016/j.patcog.2017.02.025.
- [21] P. Zhang, G. Liu, and J. Song, "MFSJMI: Multi-label feature selection considering join mutual information and interaction weight," *Pattern Recognit.*, vol. 138, p. 109378, Jun. 2023, doi: 10.1016/j.patcog.2023.109378.
- [22] M. Hatami, S. R. Mahmood, and P. Moradi, "A Graph-based Multi-Label Feature Selection using ant Colony Optimization," in *2020 10th International Symposium on Telecommunications (IST)*, Dec. 2020, pp. 175–180. doi: 10.1109/IST50524.2020.9345913.
- [23] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948, doi: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- [24] S. R. Mahmood, M. Hatami, and P. Moradi, "A Trust-based Recommender System by Integration of Graph Clustering and Ant Colony Optimization," in *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, Oct. 2020, pp. 598–604. doi: 10.1109/ICCKE50421.2020.9303647.
- [25] X.-Z. Wu and Z.-H. Zhou, "A Unified View of Multi-Label Performance Measures." arXiv, Sep. 01, 2017. doi: 10.48550/arXiv.1609.00288.