© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm
e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

Examining Swarm Intelligence-based Feature Selection for Multi-Label Classification

Awder Mohammed Ahmed[1*], Adnan Mohsin Abdulazeez[2]

[1]Duhok Polytechnic University, Duhok, Kurdistan Region-IRAQ

[2]Duhok Polytechnic University, Duhok, Kurdistan Region-IRAQ

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2021.02.02.006
Received 13 February 2021; Accepted 01 July 2021; Available online 15 October 2021

Abstract: Multi-label classification addresses the issues that more than one class label assigns to each instance. Many real-world multi-label classification tasks are high-dimensional due to digital technologies, leading to reduced performance of traditional multi-label classifiers. Feature selection is a common and successful approach to tackling this problem by retaining relevant features and eliminating redundant ones to reduce dimensionality. There is several feature selection that is successfully applied in multi-label learning. Most of those features are wrapper methods that employ a multi-label classifier in their processes. They run a classifier in each step, which requires a high computational cost, and thus they suffer from scalability issues. Filter methods are introduced to evaluate the feature subsets using information-theoretic mechanisms instead of running classifiers to deal with this issue. Most of the existing researches and review papers dealing with feature selection in single-label data. While, recently multi-label classification has a wide range of real-world applications such as image classification, emotion analysis, text mining, and bioinformatics. Moreover, researchers have recently focused on applying swarm intelligence methods in selecting prominent features of multi-label data. To the best of our knowledge, there is no review paper that reviews swarm intelligence-based methods for multi-label feature selection. Thus, in this paper, we provide a comprehensive review of different swarm intelligence and evolutionary computing methods of feature selection presented for multi-label classification tasks. To this end, in this review, we have investigated most of the well-known and state-of-the-art methods and categorize them based on different perspectives. We then provided the main characteristics of the existing multi-label feature selection techniques and compared them analytically. We also introduce benchmarks, evaluation measures, and standard datasets to facilitate research in this field. Moreover, we performed some experiments to compare existing works, and at the end of this survey, some challenges, issues, and open problems of this field are introduced to be considered by researchers in the future.

Keywords: Multi-Label classification, feature selection, wrapper methods, filter methods, curse of the dimensionality

1. Introduction

Data mining is a process of finding useful knowledge from large data sets using machine learning and statistical methods. Data mining aims to extract useful patterns from data[1],[2]. We have recently been faced with the rise of big data, which consists of numerous features, due to the improvement of Internet technologies such as social networks. For example, there are thousands of distinct words in a collection of text documents, which is regarded as high-dimensional data. These types of data affect the performance of classifiers. To improve the accuracy of the learning models, the dimensionality of data needs to be decreased. A frequently used approach to address this problem is Feature Selection (FS), which eliminates redundant and irrelevant characteristics. FS aims to reduce the input space and select a small subset of features. The wrapper, filter, and embedded approaches are classified into FS methods. In contrast, wrapper methods first search through the search space and then use a learning model to evaluate sub-sets of features; filter methods use theoretical information measures to evaluate a subset of features. In comparison to filter

*Corresponding author: awder.ahmed@spu.edu.iq

methods, wrapper methods have better performance, but they rely on a classifier that requires a high computational cost by combining the choice of functions as part of a learning model. Embedded forms use the advantage of both wrapper and filter approaches.

Due to interacting with the learning model, the methods of this category have better performance than filter methods and are better than the wrapper approaches because they do not need to evaluate the feature subsets iteratively. Among them, filter methods do not require high computational resources, and they are much more scalable than the others. Filter techniques do not require any model of learning and are therefore much faster than the wrapper methods. This is why they have recently been used for many applications in the real world. On the other hand, in many real-world applications such as text classification[3], image annotation [4],[5], gene selection [6], information retrieval [7], face recognition [8],[9], and cancer classification [10],[11], each instance can be associated with multiple class labels. In these tasks, more than one class label can be assigned to each instance. Methods of feature choice are also widely applied to the tasks of multi-label classification. The multi-label FS methods are categorized into transformation and adaptive. The first category (such as Binary relevance (BP)[1], label power set (LP), and pruned problem transformation (PPT)) aims at transforming the multi-label classification task into a single-label task and then apply single-label FS methods to identify prominent features. While adaptive methods (such as MDMR [12],[13], MUMI [14], and MCLS [15]) search directly through the whole solution space to identify the feature set.

Based on previous studies, it is evident that filter-based techniques are successful than others. In a way, it is possible to classify the filter-based techniques into two categories: univariate and multivariate strategies. The first category (univariate methods) only considers the relevance between the characteristics and labels defined by an information-theoretical criterion and ignores the redundancy of the characteristics. They are thus unable to define redundant features. A number of univariate measures in the literature have been proposed, including chi-square, Information Gain (IG), F-score [16], multi-label feature selection algorithm (MLFS) [17], a multi-label memetic feature selection for text categorization using label frequency difference called (LFD) [10], and a fast algorithm called (MGFS) [19] are some of the methods in this category. On the other hand, multivariate methods consider both the correlation between selected features and their relevance to the set of labels. There are some multivariate methods, such as a MUMI [14], MDMR [12], IGMF [19], SCLS [20], and MLACO [21].

Most feature selection techniques deal with single-label data, where instances are assigned by only a single-label [22]. Some review papers, such as [23-25], introduce and compare existing single-label feature selection methods. In contrast, most of many real-world applications are multi-label classification tasks where each instance is associated with multiple class labels. Feature selection methods are also widely applied to multi-label classification tasks. Unlike single-label feature selection methods, in the multi-label feature selection tasks, the relation among features and several labels should be considered in the selection task, which is a complicated process. There are many challenges in working with high-dimensional data in multi-label learning. Plenty of feature selection methods for multi-label data are introduced to solve these issues. While there are only a few papers that introduce and compare the existing works in this area. For example, in [26], only four data transformation methods are evaluated. This work does not consider adaptive methods, which is a majority part of multi-label feature selection methods. Moreover, the authors of [27] first proposed a multi-label feature selection method called LCFS, which considers the correlation of labels in its process. Moreover, as another contribution, they review existing single and multi-label feature selection methods. However, they only provide a taxonomy of existing multi-label feature selection methods, and they ignore comparing and evaluating existing methods in their review process. More recently, [28] provides a good categorization of existing multi-label feature selection methods. To the best of our knowledge, there is no more review paper in this area. Despite some review papers, there is still a need to provide a review of existing works in an analytical manner. Moreover, in recent years researchers have focused on applying swarm intelligence and evolutionary computing methods such as Particle Swarm Optimisation (PSO), Ant Colony Optimisation (ACO), Artificial Bee Colony (ABC), and Genetic Algorithm (GA) in selecting features of multi-label classification tasks. In [24], the authors recently provided a comprehensive review of the single-label-based feature selection method without paying attention to the existing multi-label feature selection method. To the best of our knowledge, there is no review paper to consider the existing swarm intelligence-based feature selection methods. Thus, this paper provides a systematic review of various swarm intelligence and evolutionary computing methods of feature selection presented for multi-label classification tasks. To this end, in this review, we have investigated most of the well-known and state-of-the-art methods and categorize them based on different perspectives. The main properties of the existing multi-label feature selection methods are introduced. Thereafter, some experiments are performed on these methods to compare them in an analytical manner. We also introduce benchmarks, evaluation measures, and standard datasets to facilitate research in this field. Moreover, we performed some experiments to compare existing works, and at the end of this survey, some challenges, issues, and open problems of this field are introduced to be considered by researchers in the future.

The remainder of the paper is organized as follows. Section 2 provides the formal definition of multi-label classification and the category of existing methods for classification multi-label data. Section 3 presents and reviews existing multi-label feature selection methods. Section 4 categorizes the existing methods and summarizes them based on their main properties. Section 4 discusses the experimental setup of multi-label feature selection methods, including

the evaluation measures, frequently used and real-world datasets, and also reports the comparison results obtained on the performed experiments on the existing works. Finally, the paper is concluded in Section 5.

2. Multi-label Learning

Multi-label classification, in machine learning, referred to a set of tasks with multi-output classification. It is a variant of supervised learning, where each sample is assigned multiple labels. Many real-world applications are in semantic relationships with more than one label, such as text classification, image annotation, protein function classification, music categorization, and semantic scene classification. For example, one can assign more than one label (i.e., Sky, Man, Dog, Shoes, and so on) to the following image. Fig. 1 shows a multi-label classification example.

Fig. 1 - Multi-label classification example

As another example, in the text classification task, each document can simultaneously belong to more than one topic. We first demonstrate the formal definition of the multi-label data, and then we provide a taxonomy of the multi-label classification algorithms. Suppose $X = R^M$ denotes an M-dimensional sample space and $L = \{y_1, y_2, \dots, y_q\}$ is the label space with q distinct labels. The multi-label classification problem is to train a function $h: X \to 2^q$ from the training set $TS = \{(x_i, Y_i), i = 1,2, \dots, N\}$ with N samples that each sample is a vector in the form of $x_i = \{x_{i1}, x_{i2}, \dots, x_{iM}\}$ presented by M features and a set of q labels $Y_i = \{y_{i1}, y_{i2}, \dots, y_{iq}\}$. This representation is shown in Fig. 2:

X				Y			
				y_1	y_2	$\cdots$	y_q
X_{11}	X_{12}	$\cdots$	X_{1M}	0	1	1	0
X_{21}	X_{22}	$\cdots$	X_{2M}	1	1	0	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{N1}	X_{N2}	$\cdots$	X_{NM}	1	0	1	0

Fig. 2 - Multi-label instance representation

Till now, many methods have been proposed to solve the multi-label task from various perspectives and have gained valuable results. The training set consists of samples assigned by a set of labels in the multi-label classification, and the task is to predict the label sets of unseen instances by analyzing training instances with known label sets. Multi-label classification methods are generally categorized into (1) problem conversion and (2) adaptive methods. Additional details regarding these methods are described in the following subsections.

2.1 Transformation Methods

The objective of problem transformation methods is to transform the task into a single-label issue and then apply traditional classifiers such as Neural Networks or Naïve Bayes. This approach can be carried out through three primary methods: Binary relevance, Classifier chains, and Label power set. On the other hand, transformation methods aim to map the multi-label classification of the problem into several single-label tasks. Afterward, to decrease the

dimensionality, any single-label feature selection techniques can be used. Binary Relevance (BR) (Boutell et al., 2004) aims at converting the multi-label into several binary classification problems: one for each label. The label power set (LP) method [29] considers each unique subset of labels as a single-label and uses a single-label classifier. Using a single-label classifier to classify an unseen sample, this method predicts its associated label set and then maps it back to the power set. In this method, the number of classes grows by increasing the labels, although it considers the correlation of the labels. However, predictive performance and the computational efficiency of LP are reduced with the growing number of labels and training samples. The PPT (Pruned Problem Transformation) method is an extension of the LP process. The concept of PPT is to prune those sets of labels that have appeared below a predefined threshold. Instances belonging to the pruned labels can either be assigned or eliminated to another class. This strategy reduces the complexity and classification error associated with a large number of infrequent sets. The entropy-based label assignment method (ELA) uses labels' entropy to assign labels' weights in a multi-label document.

2.2 Adaptive Methods

Adaptive methods directly perform multi-label classification without transforming into single-label classification tasks. To define the final set of features, adaptive methods scan directly through the entire solution space of features and labels. For example, MUMI [14] maximizes the correlation between chosen features and the class labels. This method uses the concept of mutual information in an incremental way to compute the correlation between variables. [12] incorporates the parameters for max-dependency and min-redundancy with reciprocal knowledge for the gradual estimation of class label features. The dependence shows how a function relates to each label, and the redundancy shows the connection between the candidate features and the entire labels considered by the selected features. [30] proposed a method that considers the correlation between features and labels without converting the task into a single-label problem. [31] proposed three measures based on mutual neighborhood information to compute the quality of candidate features. These measures use the margin of the instance to granulate all instances under various labels.
The remainder of this paper is organized as follows. Section Two describes some researches works that are listed and discussed. In section three, all mentioned and reviewed researches are compared. Finally, in section four, the conclusion of this work is presented.

3. Related Work

The main objective of the selection of multi-label features is to identify a set of prominent characteristics with minimum redundancies that are most relevant to a set of target classes. In general, these methods are categorized into the transformation of data and adaptive techniques Methods of transformation aim to convert the multi-label task to a single-label study and then use a traditional feature selection method to decrease dimensionality. In contrast, adoption methods are applied directly to the multi-label space to reduce dimensionality. A majority part of single-table and multi-label feature selection methods employ swarm intelligence methods. To this end, in this section, we have a comprehensive survey on swarm intelligence-based feature selection methods, and then we have reviewed existing multi-label transformation and adoption methods. In their corresponding sections, the details of these methods are described.

For feature selection, swarm intelligence such as Particle Swarm Optimisation (PSO)[32],[33],[34], Artificial Bee Colony (ABC)[35-37], and ACO and evolutionary methods such as Genetic Algorithm (GA) [30] have been used successfully. Among them, ACO has gained better results than the others. In general, ACO-based methods are categorized into Wrapper and Filter methods. To evaluate a set of characteristics selected by each ant through its traverse, wrapper techniques use a learning model. Filter methods, by contrast, use an information-theoretical measure to evaluate the results of ants. For example, the authors of [38,39] proposed an ACO-based unsupervised feature selection method called UFSACO. This method uses a measure to compute the redundancy of features. The relevance of features to the target class is not considered by this method, however.

The same authors suggested a multivariate filter-based relevance-redundancy feature selection method called RRFSACO in [38]. This method uses ACO and considers in its search process both the concepts of relevance and redundancy. In [26], an unsupervised probabilistic feature selection called UPFS is proposed. In order to reduce the redundancy between functions within the iterative search process of ACO, this method uses ACO and looks for the optimal feature subset by considering the inter-feature data. The authors [40] proposed an unsupervised filter method called GCACO, which combines ACO with graph clustering for feature selection. GCACO first divides similar features into clusters and then uses the search strategy of ACO to rank features. This method reduces the redundancy by force the method to prevent choosing similar features. To this aim, GCACO assigns a penalty for ants to remain in the same cluster; the authors of [41] suggested a method called MGCACO, which extends GCACO by proposing a measure to evaluate both the relevance and redundancy of characteristics using multiple discriminant analysis (MDA). [42] proposed an ACO-based feature selection ABACO which combines Artificial Bee Colony (ABC) with ACO. This method's features are mapped to a fully connected graph where each node has two sub-nodes for determining the selection or deselection of features. Ant colony optimization algorithms have been used to select nodes [30] proposed a modified binary-coded ACO integrated with the genetic algorithm. [43] proposed the method called MBACO uses the

pheromone density model to initialize pheromones. Recently [44] proposed a specific update rule for ACO, which prevents the algorithm from falling into the local optima. To this aim, this method improves the path transfer probability method by adding pheromones to more paths. The feature selection task is regarded as a multi-objective task by some ACO-based methods. For example, [45] suggested a multi-objective ACO-based feature selection method to classify disturbances in power quality. This method evaluates the solutions by using two contradictory objectives: reducing the feature subset's size and improving the classifier performance. The authors of [46] introduced a memory to keep the best ants and heuristic desirability in ACO. It is computed using a specific strategy to make the algorithm computationally efficient than its ancestors.

The authors of [47] proposed the key component analysis (PCA) is used to classify redundant characteristics and then to pick the final function subset using the genetic algorithm(GA). This method uses the multi-label naïve Bayes (MLNB) classified to evaluate each solution of GA. The authors of [17] proposed a classification algorithm for multi-label features (through the integration of shared knowledge with GA. This method uses GA to search through the solution space and employ mutual information to determine the importance of the features for each label. The authors of [31] use mutual information to compute the label granularity to identify the relationships between labels and features. In [48], constrained convex optimization was used to maximize relevance and minimizing redundancy. Simultaneously proposed a measure called "label frequency difference" (LFD) to compute the conditional frequencies of labels to recognize the discrimination power of features. LRFS [49] analyzed the differences between labels and them into two independent and dependent groups on computing the label redundancy.

A PSO-based multi-objective multi-label feature choice algorithm called MPSOFS was suggested in [50]. This method first transforms the feature selection task to an ongoing problem. It employs the ideas of non-dominated comparison and the crowding distance a probability-based encoding to transform a discrete feature selection problem into a continuous one suitable. The authors employed the idea to prune the archive. IGMF [19] computes the label correlations by using the Information gain between features and labels. Recently, the authors of [51] presented a many-objective optimization-based multi-label feature selection algorithm (MMFS). Moreover, in [34], a PSO-based multi-label learning and the arrival of features in an online fashion was proposed. The authors of [52] proposed a multi-label and filter-based feature selection method by using the search process of ACO. This method uses mutual information to evaluate solutions founded by the ants, and thus it is classified as a filter method.

4. Comparison and Discussion

In section two, each transformation method and an adaptive method has been viewed. Recently, researchers have focused on adaptive and swarm intelligence-based methods due to their success in gaining performance. This is why, in this section, we have only compared these methods. Table 1 summarizes the main properties of the existing works.

Multi-label feature selection methods can be classified according to the perspective of label and search strategy. Considering the label perspective, training samples can be labeled, unlabelled, and semi-labeled. Therefore, this property leads to the prescience of three categories of multi-label feature selection methods include supervised, unsupervised, and semi-supervised feature selection methods, respectively. It is assumed that training samples contain class labels in supervised methods and are expected to have high accuracy. Multi-label ranker (MLFR)[53], multi-label feature selection based on the information gain (IGMF)[54], and multi-label correlation-based feature selection (ML-CFS)[55] are some well-known examples in this category. The presence of the labels in the training samples leads to enhance the accuracy. On the other hand, all wrapper methods such as MR2PSO [10], ABACO [13], OFS [20], and MPSOFS [26] require the class labels to train the multi-label classifiers. Moreover, the wrapper methods in each iteration use a learning model such as MLKNN or MLNB to evaluate the feature subset. Training a learning model is a time-consuming process, and thus, we can claim that wrapper methods can be scalable for large-scale and real-world applications. On the other hand, the filter methods employ some information-theoretic criteria to evaluate feature subsets instead of running a learning model. Therefore, these methods are much faster than the wrapper methods. For example, MGCACO[16], MDMR[3], MUMI[5], MICO [45], and FACO [19] are some well-known supervised filter methods. All these methods require the class labels to search through the solution space. While in many real-world applications, it is hard or time-consuming to provide labels to the training samples.

Table 1 - Summarizes the main properties of the existing works

Methods	Year	Single label Multi-label	Search Process	Approach type	Relevancy-Redundancy	Filter\Wrapper
MMLACO[49]	2021	ML	ACO	AD	MV	Filter
MMOFS[31]	2021	ML	PSO	AD	MV	Filter
MMFS[48]	2020	ML	GA	AD	MV	Wrapper
FMABC-FS[34]	2020	SL	ABC	-	-	Filter

MICO [45]	2019	ML	MI	AD	MV	Filter
LRFS [29]	2019	ML	MI	AD	MV	Filter
MLACO[20]	2019	ML	ACO	AD	MV	Filter
OFS[20]	2019	SL	ACO	-	-	Wrapper
WFACOFS [43]	2019	SL	ACO	-	-	Wrapper - Filter
MGCACO[16]	2018	SL	ACO	-	-	Filter
FACO [19]	2018	SL	ACO	-	-	Filter
SCLS[25]	2017	ML	MI	AD	MV	Filter
MPSOFS [26]	2017	ML	PSO	AD	UV	Wrapper
GMFS[27]	2017	ML	MI	AD	MV	Wrapper
MFNMI[24]	2016	ML	MI	AD	MV	Filter
MBACO[14]	2016	SL	ACO + GA	-	-	Wrapper
MDMR[3]	2015	ML	MI	AD	MV	Filter
GA-ML-CFS[23]	2015	ML	GA	AD	UV	Filter
ABACO[13]	2015	SL	ACO	-	-	Wrapper
RRFSACO[12]	2015	SL	ACO	-	-	Filter
GCACO[8]	2015	SL	ACO	-	-	Filter
PSO-RR[11]	2014	SL	PSO	-	-	Wrapper
UFSACO[21]	2014	SL	ACO	-	-	Filter
MUMI[5]	2013	ML	MI	AD	MV	Filter
MR2PSO[10]	2011	SL	PSO	-	-	Wrapper

Table 1 - Some similar methods of multi-label function collection. DT: Data Transformation, AD: Adaptive Methods, IG: Information Gain, ACO: Ant Colony Optimization, GA: Genetic Algorithm, NB: Naïve Bayes, MI: Mutual Information, PR: Page Rank UV: Univariate, MV: Multivariate, SL: Single Label, ML: Multi-Label, DT: Data Transformation, MI: Mutual Information, PR: Page Rank UV: Univariate, MV: Multivariate

Unsupervised feature selection methods are categorized as filter methods, which do not require any class labels. MR2PSO [10] are well-known and fast unsupervised filter methods whose performance is comparable with many wrappers and supervised filter methods. This method uses the ACO in its search process and employs a specific metric to evaluate the ants' features. Moreover, there are many samples without class labels in many real-world applications, and only a few samples consist of class labels. In this situation, supervised learning methods cannot be utilized because a few samples contain class labels. Therefore, it is crucial to develop some semi-supervised feature selection methods, which consider both labeled and unlabelled training samples in their processes. A majority of semi-supervised feature selection methods convert the solution space into a graph and then propagate the labels of those labeled samples to unlabelled ones.

5. Experimental Evaluations

In this section, to evaluate the existing woks' performance, it is performed on six well-known real-world and diverse datasets. These datasets include Arts, Birds, CAL500, Computer, Corel5K, Education, and Yeast. All these datasets can be accessed from Mulan Library1. Table 2 summarizes the properties of these datasets. This table for each dataset shows the dataset name (Name), the number of samples (Instance), the number of features (Features), the number of labels (Labels), dataset density (Density), feature type (type), and dataset domain (Domain).

Table 2 - Description of multi-label datasets

Dataset	Instances	Features	Labels	Density	type	Domain
Arts	5000	462	26	0.063	Numeric	Text
Birds	645	260	19	0.053	Numeric	Audio
CAL500	502	68	174	0.150	Numeric	Music

[1] http://mulan.sourceforge.net/datasets-mlc.html

Corel5K	5000	499	374	0.009	Nominal	Image
Education	5000	550	33	0.044	Numeric	Text
Yeast	2417	103	14	0.303	Numeric	Biology

Moreover, our method is compared to six state-of-art multi-label feature selection, and filter-based methods include MDMR [12], SCLS [20], MGFS [19], MLACO [21], AMI [53], and IGMF [19]. In the experiments, in order to evaluate the performance of the proposed method, the ML-kNN classifier has been employed. The ML-kNN is a multi-label version of the traditional and famous kNN (k-nearest neighbor classifier) [56] with k=10 neighbors. ML-kNN works by first detecting the k nearest neighbors for each sample in training data, and then the test samples will be assigned to the labels set, which is most popular among its neighbors. Actually, it identifies k nearest neighbors for each unseen instance in the training set. Then, the maximum a posteriori (MAP), statistical information gained from the label set, has been applied to determine the set of labels for the unseen instances. In single-label feature selection, an instance can be classified incorrectly or correctly. While, in multi-label feature selection, the problem is much harder and difficult than single-label feature selection, as the predicted label subset is very different from the actual label subset. Thus, there are different criteria evaluation for multi-label methods than those used in single-labels. To show the power of prediction of the used classifier ML-kNN and then show our method's performance, we have used five well-known multi-label evaluation measures: Hamming loss, Average precision, Ranking loss, F1-micro, and F1-macro [57-59].

Hamming loss: This measure computes the average difference between the predicted and ground-truth labels, and it is defined as:

$$HL = \frac{1}{N}\sum_{i=1}^{N}\frac{y'_i \oplus y_i}{m} \tag{1}$$

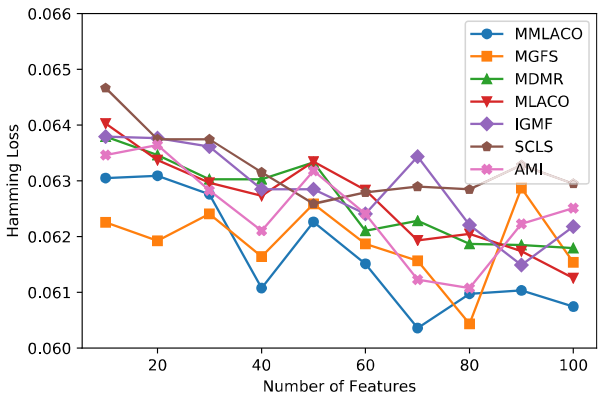
Hamming loss will be achieved to the best performance when it is approaching 0.

Ranking Loss: This measure evaluates the number of occurrences that relevant labels are ranked lower than irrelevant labels.

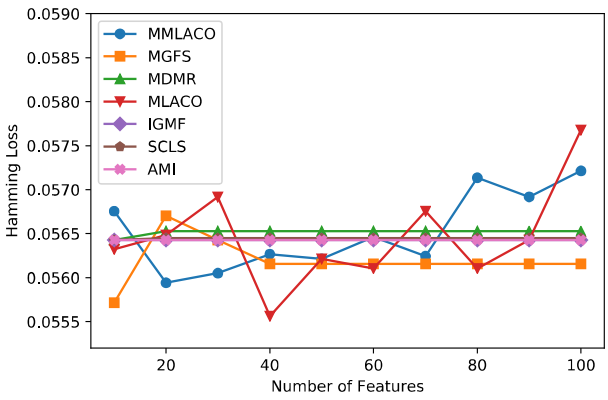
$$RL = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i||\bar{y}_i|}\left|(\lambda_1,\lambda_2)\right| \lambda_1 \le \lambda_2, (\lambda_1,\lambda_2) \in y_i * \bar{y}_i \tag{2}$$

where λ_j shows a likelihood between $x_i$and each label$l_i \in L$, and $\bar{y}_i$ is the complementary set of y_i. The smaller value of it shows better performance.

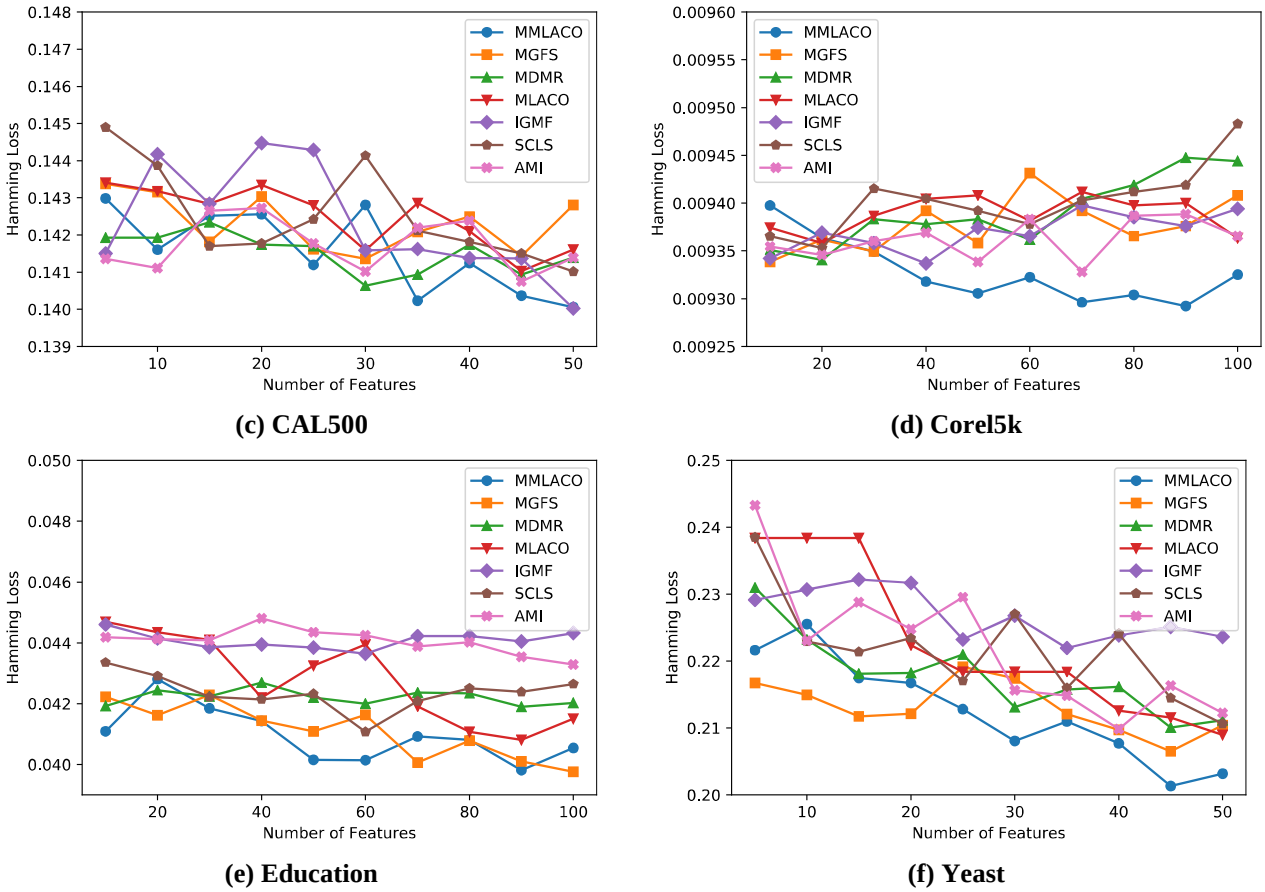
The methods are evaluated using a various number of features, and the results are reported in Figs. 3 (a) - (f). The vertical axis of these figures indicates the classification performance value in terms of hamming loss and ranking loss measures, respectively. The results are reported over five independent runs. It is clear from Fig. 3 that ACO-based methods achieved smaller hamming loss criteria. This value is significantly increased while a higher number of features are chosen. Moreover, most multi-label filter-based feature selection methods do not consider both relevancy and redundancy. Some of them only consider the count of features without paying attention to the ants that choose the feature. Thus, these methods are not capable of finding and removing redundant features as well. To tackle this issue, some methods such as MMLACO are able of find and eliminate redundant and irrelevant features.



(a) Arts



(b) Birds

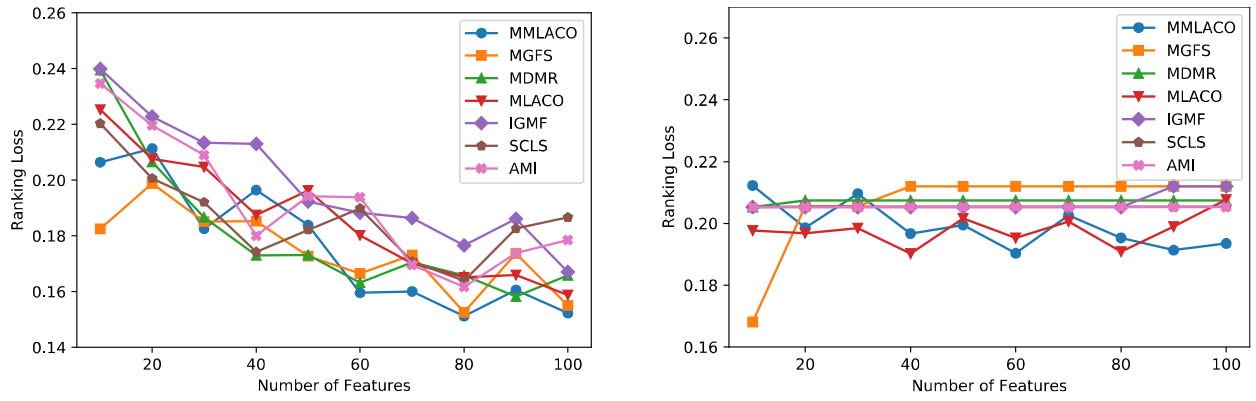


(c) CAL500 **(d) Corel5k**

(e) Education **(f) Yeast**

Fig. 3 - The Hamming loss values of the multi-label feature selection methods with the various number of features

Moreover, similar results were achieved where the experiments were reported using the ranking loss measure. It is clear from Fig. 4 (a) - (f) that those methods which consider both relevancy and redundancy concepts in their search processes could achieve better Ranking loss values. Moreover, from the results, it can be concluded that swarm intelligence-based methods such as MMLACO and MLACO could better search through the search space and find more relevant features compared to the heuristic-based methods such as MGFS and MDMR.



(a) Arts **(b) Birds**

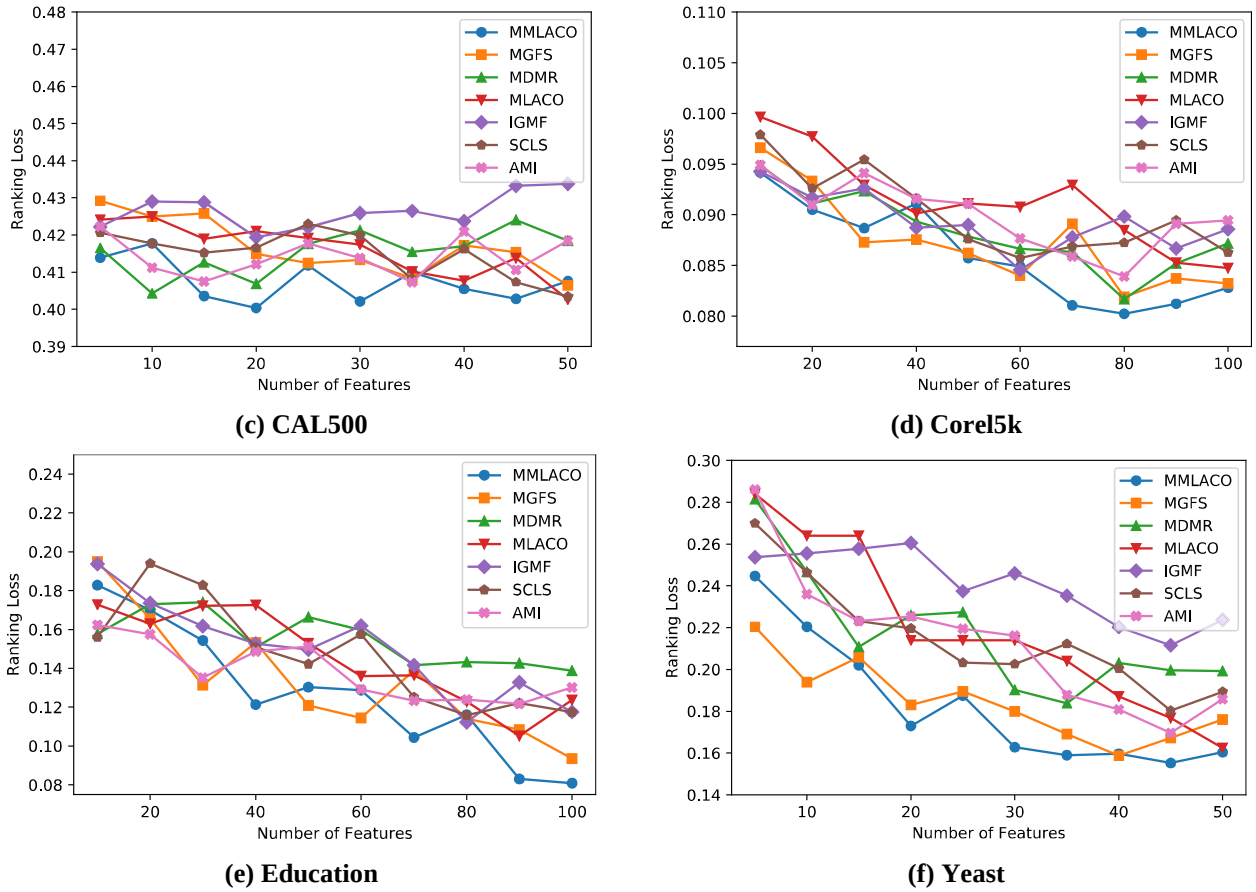


Fig. 4 - The Ranking loss values of the multi-label feature selection methods with the various number of features

6. Conclusion

In this work, we presented a detailed survey on multi-label feature selection methods based on swarm intelligence. Such techniques are categorized into wrapper and filter methods. To test the feature set, wrapper methods use a multi-label classifier, such as MLKNN or MLNB. These methods yield precise results; however, they require a high computing cost to classify a range of prominent features due to the classifier's use. Simultaneously, filter methods use an information-theoretical criterion to measure the similarity between features and compute the significance of each feature with labels, such as shared information. They are, thus, much quicker than the wrapper techniques. On the other hand, most of these approaches use a swarm intelligence method such as PSO, ACO, ABC, and Genetic Algorithm to search for a feature set across the solution space. We have conducted an analysis in this paper to compare the current methods of selection of multi-label features. The findings show that compared to other swarm-intelligence approaches, ACO-based methods yield better results. A potential solution for enhancing swarm-intelligence methods' search process is by clustering the search space and related community features into a single cluster.

Acknowledgment

The authors would like to acknowledge Duhok Polytechnic University/ Technical College of Informatics- Akre.

References

[1] Abdulazeez, A. M., Sulaiman, M. A. and Zeebaree, D. Q. (2020). Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. Journal of Soft Computing and Data Mining, 1(2), 11-25

[2] Mostafa, S. A., Mustapha, A., Mohammed, M. A., Ahmad, M. S., & Mahmoud, M. A. (2018). A fuzzy logic control in adjustable autonomy of a multi-agent system for an automated elderly movement monitoring application. *International journal of medical informatics*, *112*, 173-184

[3] Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. Knowledge-Based Systems, 163, 955-971

[4] Zhu, P., Xu, Q., Hu, Q., Zhang, C., & Zhao, H. (2018). Multi-label feature selection with missing labels. Pattern Recognition, 74, 488-502
[5] Zhang, Y., Wu, J., Cai, Z., & Philip, S. Y. (2020). Multi-view multi-label learning with sparse feature selection for image annotation. IEEE Transactions on Multimedia, 22(11), 2844-2857
[6] Tabakhi, S., Najafi, A., Ranjbar, R., & Moradi, P. (2015). Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing, 168, 1024-1036
[7] Kadhe, S., Garcia, B., Heidarzadeh, A., El Rouayheb, S., & Sprintson, A. (2019). Private information retrieval with side information. IEEE Transactions on Information Theory, 66(4), 2032-2043
[8] Kute, R. S., Vyas, V., & Anuse, A. (2019). Component-based face recognition under transfer learning for forensic applications. Information Sciences, 476, 176-191
[9] Du, H., Ma, L., Li, G., & Wang, S. (2020). Low-rank graph preserving discriminative dictionary learning for image recognition. Knowledge-Based Systems, 187, 104823
[10] Jain, I., Jain, V. K., & Jain, R. (2018). Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing, 62, 203-215
[11] Ting, F. F., Tan, Y. J., & Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications, 120, 103-115
[12] Almugren, N., & Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access, 7, 78533-78548. [13] Lin, Y., Hu, Q., Liu, J., & Duan, J. (2015). Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing, 168, 92-103
[14] Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved threshold based and trainable fully automated segmentation for breast cancer boundary and pectoral muscle in mammogram images. Ieee Access, 8, 203097-203116
[15] Lee, J., & Kim, D. W. (2013). Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters, 34(3), 349-357
[16] Huang, R., Jiang, W., & Sun, G. (2018). Manifold-based constraint Laplacian score for multi-label feature selection. Pattern Recognition Letters, 112, 346-352
[17] Ding, S. (2009). Feature selection based F-score and ACO algorithm in support vector machine. In 2009 Second International Symposium on Knowledge Acquisition and Modeling (Vol. 1, pp. 19-23). IEEE
[18] Yu, Y., & Wang, Y. (2014). Feature selection for multi-label learning using mutual information and GA. In International Conference on Rough Sets and Knowledge Technology (pp. 454-463). Springer, Cham
[19] Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-pour, H. (2020). MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality. Expert Systems with Applications, 142, 113024
[20] Li, L., Liu, H., Ma, Z., Mo, Y., Duan, Z., Zhou, J., & Zhao, J. (2014). Multi-label feature selection via information gain. In International Conference on Advanced Data Mining and Applications (pp. 345-355). Springer, Cham
[21] Lee, J., & Kim, D. W. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. Pattern Recognition, 66, 342-352
[22] Mostafa, S. A., Mustapha, A., Mohammed, M. A., Hamed, R. I., Arunkumar, N., Abd Ghani, M. K., ... & Khaleefah, S. H. (2019). Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cognitive Systems Research, 54, 90-99
[23] Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019). Trainable model based on new uniform LBP feature to identify the risk of the breast cancer. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 106-111). IEEE
[24] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375
[25] Rostami, M., Berahmand, K., Nasiri, E., & Forouzande, S. (2021). Review of swarm intelligence-based feature selection methods. Engineering Applications of Artificial Intelligence, 100, 104210
[26] Di Mauro, M., Galatro, G., Fortino, G., & Liotta, A. (2021). Supervised feature selection techniques in network intrusion detection: A critical review. Engineering Applications of Artificial Intelligence, 101, 104216
[27] Spolaôr, N., Cherman, E. A., Monard, M. C., & Lee, H. D. (2013). A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science, 292, 135-151
[28] Spolaôr, N., Monard, M. C., Tsoumakas, G., & Lee, H. D. (2016). A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing, 180, 3-15
[29] Kashef, S., Nezamabadi‐pour, H., & Nikpour, B. (2018). Multi-label feature selection: A comprehensive review and guiding experiments. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(2), e1240.
[30] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Random k-labelsets for multi-label classification. IEEE transactions on knowledge and data engineering, 23(7), 1079-1089
[31] Jungjit, S., & Freitas, A. A. (2015, April). A new genetic algorithm for multi-label correlation-based feature selection. In 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine

Learning (pp. 285-290). [32] Li, F., Miao, D., & Pedrycz, W. (2017). Granular multi-label feature selection based on mutual information. Pattern Recognition, 67, 410-423
[33] Moradi, P., & Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing, 43, 117-130
[34] Zebari, D. A., Zeebaree, D. Q., Saeed, J. N., Zebari, N. A., & Adel, A. Z. (2020). Image steganography based on swarm intelligence algorithms: A survey. people, 7(8), 9
[35] Paul, D., Jain, A., Saha, S., & Mathew, J. (2021). Multi-objective PSO based online feature selection for multi-label classification. Knowledge-Based Systems, 222, 106966
[36] Zhang, Y., Cheng, S., Shi, Y., Gong, D. W., & Zhao, X. (2019). Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm. Expert Systems with Applications, 137, 46-58. [37] Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., ... & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing, 74, 634-642
[38] Wang, X. H., Zhang, Y., Sun, X. Y., Wang, Y. L., & Du, C. H. (2020). Multi-objective feature selection based on artificial bee colony: An acceleration approach with variable sample size. Applied Soft Computing, 88, 106041.
[39] Tabakhi, S., & Moradi, P. (2015). Relevance–redundancy feature selection based on ant colony optimization. Pattern recognition, 48(9), 2798-2811
[40] Tabakhi, S., Moradi, P., & Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence, 32, 112-123
[41] Moradi, P., & Rostami, M. (2015). Integration of graph clustering with ant colony optimization for feature selection. Knowledge-Based Systems, 84, 144-161
[42] Ghimatgar, H., Kazemi, K., Helfroush, M. S., & Aarabi, A. (2018). An improved feature selection algorithm based on graph clustering and ant colony optimization. Knowledge-Based Systems, 159, 270-285. [43
[43] Kashef, S., & Nezamabadi-pour, H. (2015). An advanced ACO algorithm for feature subset selection. Neurocomputing, 147, 271-279
[44] Wan, Y., Wang, M., Ye, Z., & Lai, X. (2016). A feature selection method based on modified binary coded ant colony optimization algorithm. Applied Soft Computing, 49, 248-258
[45] Peng, H., Ying, C., Tan, S., Hu, B., & Sun, Z. (2018). An improved feature selection algorithm based on ant colony optimization. IEEE Access, 6, 69203-69209
[46] Singh, U., & Singh, S. N. (2019). A new optimal feature selection scheme for classification of power quality disturbances based on ant colony framework. Applied Soft Computing, 74, 216-225
[47] Manosij, G., Ritam, G., Sarkar, R., & Abraham, A. (2020). A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing & Applications, 32(12), 7839-7857
[48] Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. Information Sciences, 179(19), 3218-3229
[49] Sun, Z., Zhang, J., Dai, L., Li, C., Zhou, C., Xin, J., & Li, S. (2019). Mutual information based multi-label feature selection via constrained convex optimization. Neurocomputing, 329, 447-456
[50] Zhang, P., Liu, G., & Gao, W. (2019). Distinguishing two types of labels for multi-label feature selection. Pattern Recognition, 95, 72-82
[51] Zhang, Y., Gong, D. W., Sun, X. Y., & Guo, Y. N. (2017). A PSO-based multi-objective multi-label feature selection method in classification. Scientific reports, 7(1), 1-12
[52] Dong, H., Sun, J., Sun, X., & Ding, R. (2020). A many-objective feature selection for multi-label classification. Knowledge-Based Systems, 208, 106456
[53] Hatami, M., Mahmood, S. R., & Moradi, P. (2020, December). A graph-based multi-label feature selection using ant colony optimization. In 2020 10th International Symposium onTelecommunications (IST) (pp. 175-180). IEEE
[54] Lastra, G., Luaces, O., Quevedo, J. R., & Bahamonde, A. (2011, October). Graphical feature selection for multi-label classification tasks. In International Symposium on Intelligent Data Analysis (pp. 246-257). Springer, Berlin, Heidelberg
[55] Li, L., Liu, H., Ma, Z., Mo, Y., Duan, Z., Zhou, J., & Zhao, J. (2014, December). Multi-label feature selection via information gain. In International Conference on Advanced Data Mining and Applications (pp. 345-355). Springer, Cham
[56] Jungjit, S., Michaelis, M., Freitas, A. A., & Cinatl, J. (2013, October). Two extensions to multi-label correlation-based feature selection: A case study in bioinformatics. In 2013 IEEE International Conference on Systems, Man, and Cybernetics (pp. 1519-1524). IEEE
[57] Lee, J., Lim, H., & Kim, D. W. (2012). Approximating mutual information for multi-label feature selection. Electronics letters, 48(15), 929-930
[58] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048
[59] Wu, X. Z., & Zhou, Z. H. (2017). A unified view of multi-label performance measures. In International Conference on Machine Learning (pp. 3780-3788). PMLR