Payame Noor University

Control and Optimization in Applied Mathematics (COAM)
Vol. 6, No. 2, Summer-Autumn 2021 (37-52), ©2016 Payame Noor University, Iran

DOI. 10.30473/coam.2021.60849.1177 (Cited this article)

Research Article

Multi-Label Classification with Meta-Label-Specific Features and Q-Learning

Seyed Hossein Seyed Ebrahimi[1], Kambiz Majidzadeh[2,*], Farhad Soleimanian Gharehchopogh[3]

[1,2,3]Department of IT and Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran.

Received: September 15, 2021; **Accepted:** January 01, 2022.

Abstract. Classification is a crucial process in data mining, data science, machine learning, and the applications of natural language processing. Classification methods distinguish the correlation between the data and the output classes. In single-label classification (SLC), each input sample is associated with only one class label. In certain real-world applications, data instances may be assigned to more than one class. The type of classification which is required in such applications is known as multi-label classification (MLC). In MLC, each sample of data is associated with a set of labels. Due to the presence of multiple class labels, the SLC learning process is not applicable to MLC tasks. Many solutions to the multi-label classification problem have been proposed, including BR, FS-DR, and LLSF. But, these methods are not as accurate as they could be. In this paper, a new multi-label classification method is proposed based on graph representation. A feature selection technique and the Q-learning method are employed to increase the accuracy of the proposed algorithm. The proposed multi-label classification algorithm is applied to various standard multi-label datasets. The results are compared with state-of-the-art algorithms based on the well-known performance evaluation metrics. Experimental results demonstrated the effectiveness of the proposed model and its superiority over the other methods.

Keywords. Machine learning, Classification, Multi-label, Meta-label specific features.

MSC. 00A15.

* Corresponding author
h.seyedebrahimi@iaurmia.ac.ir, k.majidzadeh@iaurmia.ac.ir, bonab.farhad@gmail.com

1 Introduction

Classification is one of the most important applications of statistical methods in various sciences. One of the main purposes of modeling and classification in statistics is to forecast based on the facts, available features and available information on a particular topic [1, 14, 16]. In statistics, this task is mainly the responsibility of methods such as regression, audit analysis, time series, classification, and tree regression. Classification is one of the most widely used methods in data analysis. The aim of classification methods is to extract the patterns in the data by grouping people and variables [2, 12, 13, 24]. Classification involves a wide variety of methods that are used in many sciences. These methods differ in terms of their purposes, the algorithm they use, and the way they display the results. Classification is the assignment of records or any set of objects to a specific set of categories. It has many applications in various commercial, economic and medical contexts.

Classification is one of the principles of data mining. Classification algorithms, through samples stored in the training dataset, produce general models that can be used to predict class predictions for new samples [10, 20, 22, 36]. Each instance of a training dataset is assigned to a set of attributes. Each attribute specifies a class associated with the instance. There are several classification algorithms based on the number of available classes as well as the number of classes to which each instance can belong.

In multi-label classification, each instance belongs to several classes. This type of classification originated from the field of text classification; simply because each text can belong to several predefined titles at the same time [11, 34, 9, 29]. Multi-label classification has been widely used in real-world applications such as music classification, protein function classification, photo interpretation, video classification, landscape semantic classification, and more. The purpose of multi-label classification is to create a function that maps each multi-label data to a related class set [6, 7, 28, 32, 37]. This division is shown in Figure 1.

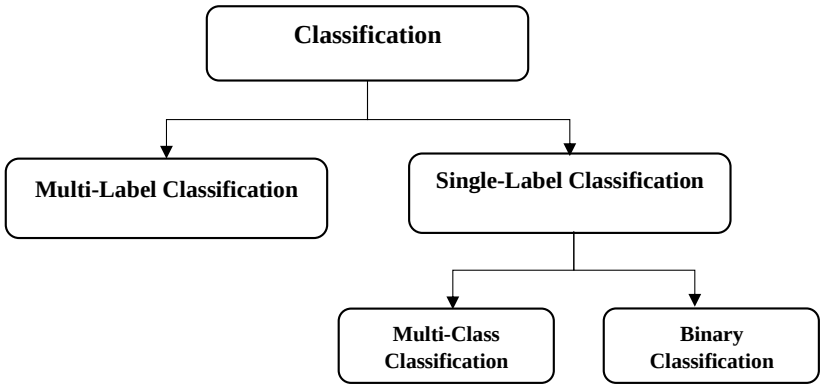

Figure 1: Categorization of the classification algorithms.

Due to the importance and widespread use of multi-label classifiers, in what follows, we introduce three important criteria that allow us to classify and compare them.

The first criterion is based on the way the learning algorithm deals with the multi-label dataset. All supervised-based learning methods, which are presented on multi-label data, can be divided into two general categories. The first group consists of *transitional problems*, and the second group involves *adaptation methods*.

A transitional problem is one that maps the problem to single-label learners, and this is why it is named so. But, the second category includes methods that can directly apply multi-label data.

Another important criterion is classification based on the output of multi-label learners. The output of any multi-label learner is either a model that classifies multiple labels or performs the label rating operation. Each test sample can be labeled with relevant and unrelated labels in a multi-label classification model. However, label ratings rank all possible labels for each sample.

In this paper, a novel classification method is presented for multi-label classification problems. In the proposed method, the Q-Learning approach and a feature selection technique are employed. Moreover, in the proposed model, the feature space is encoded as a graph, and the Q-leaning is utilized to select the most appropriate feature subset.

The main novelty of this paper is the integration of Q-learning and feature weighting to improve the accuracy of classification. In fact, this research utilizes a Q-learning-based model for final feature selection, what makes it different from previous methods.

The remainder of this paper is organized as follows. Section 2 is devoted to a review of the related works. In Section 3, the details of our proposed method are explained. The performance of the proposed method is evaluated in Section 4. Finally, the achievements of our research are outlined in a brief conclusion.

2 Related Works

In this section, we review the research literature of the field of multi-classification.

2.1 Transfer Methods

The methods in this category strive to transfer a multi-label dataset into a single-label one. This is done by breaking down the original multi-label dataset into several single-label datasets. Single-label classifiers are then applied to these datasets. Finally, all classifiers are combined to produce a multi-label classifier.

Six methods are suggested in [4] that convert any multi-label dataset into a single-label one. These methods can be described as follows.

- For each instance, select the label with the highest index, among the associated label sets.

- For each sample, select the label with the lowest index among the associated label sets.
- Copy each sample to the number of labels, and assign only one sample label to each copy.
- This is the same as the previous one, except that we assign a weight to each pair (sample, label).
- Randomly select one of the sample labels.
- Delete samples that have more than one label in the collection.

The problem with the aforementioned methods is that they destroy a lot of information involved in the initial multi-label dataset, because the samples lose some of their labels. As a result, the learning algorithm will not use all the information in the initial dataset to generate the model. It is evident that the model created in this case is less efficient than the model created using the entire initial training dataset.

In [19], a novel method called "K-way Tree-based eXtreme Multi-Label Classifier (KTXMLC)" is proposed. To maintain the correlations, this method operates on a tree-based classifier utilizing a clustering algorithm.

Moreover, in the multi-label classification problem, a novel method is developed to reweight examples in [38]. In [35], for multi-label classification, dual aggregated network is proposed on pyramidal convolutional features. To learn discriminant multi-scale information of different intended objects within the image data, this approach consists of both classifier-level aggregation and feature. In [33], even at low prior probabilities, a new label dependence criterion demonstrates values from a full range to develop a data-driven label clustering.

The method proposed in [31] describes four commonly used tricks in data analysis. Each of them converts multi-label datasets into a number of single-label ones. These methods include One by One (OBO), One Versus One (OVO), One Versus Rest (OVR), and Label Powerset (LP). Moreover, in [23] an approach based on deep learning is proposed with label-attention and domain-specific pre-training for multi-label legal document classification. In [5], an algorithm of multi-label feature selection is proposed on the basis of many-objective optimization.

In this paper, an enhanced NSGA III algorithm is employed with two archives with the aim of improving the convergence and diversity of NSGA III.

2.2 Adaptation Methods

In adaptation methods, the existing algorithms are modified or are combined with other models. Algorithms in the first category produce a special multi-label classifier that considers all instances and all classes of the training dataset at once. But the second category improves an existing single-label classifier method, while multi-label datasets are implicitly or explicitly subdivided into a sequence of subsets. Several efficient and

effective multi-label classification algorithms have been produced and designed by the second method.

Existing adaptation methods are based on the development of dependency rule learners, decision trees, sample-based methods, neural networks, cumulative methods, and SVM classifiers.

In [26], a method called MMAC is presented. This method explores the rules of dependency to create a set of classification rules. Then, it deletes the instances that follow these rules, and repeats the rule search on the remaining data. This continues until all the samples follow at least one rule. The disadvantage of this method is that it is very suitable for training samples, because the flexibility of the method on new data is reduced.

3 The Proposed Method

This section presents the details of the proposed Q-learning multi-label classification algorithm. In the proposed algorithm, the Q-learning approach is utilized to improve classification accuracy in multi-label data. Moreover, in the proposed method, a new graph-based feature selection method is introduced to reduce the dimension of data.

To describe the data, a large number of attributes are used. A majority of the so-called features may be redundant and irrelevant to the application of the intended data mining. Due to the redundant and unrelated features in the data, the performance of the machine learning algorithm may be negatively affected. Additionally, these features can also increase the complexity of computation, which is why reducing the size of dataset is one of the first steps in data mining and machine learning. The model, which is based on reduced features, has a higher generalizability than the original model. Based on a widely accepted rule, a minimum of $10 \times n \times C$ educational data is essential in order to classify a problem with dimension n and C classes. The reduction in the number of features could lead to a diminish in the amount of required training data, once it is practically impossible to present the required amount of training data. Subsequently, there would be an upsurge in the overall performance of the classification algorithm [17, 8, 18].

In higher dimensions, managing the data is difficult, and the computational and analytical capabilities reduce compared to lower dimensions. Therefore, dimensionality is an essential part of the knowledge discovery process. Multi-dimensional data platforms pose many computational challenges despite the opportunities they create. One of the problems with large datasets is that most of the time, all data features are not important for finding the knowledge that lies in the data. For this reason, in many areas, dimensionality reduction is one of the most significant issues.
The methods of dimensionality reduction are divided into two categories.

- **Feature-based extraction methods**: These methods map a multi-dimensional space into a smaller one. In fact, they form less attributes through the combination of existing attribute values, so that the so-called attributes hold all (or most) of the information included in the original attributes. These methods in turn are

divided into two categories: linear and non-linear methods. Linear methods are simpler and easier to understand and seek to find a sub-public space. The non-linear methods, which are more complex and more difficult to analyze, seek to find a subliminal subspace.

- **Methods based on feature selection:** These methods try to reduce the dimension of the data by selecting a subset of the primary features. Sometimes, data analytic techniques such as classification, work better on the reduced space. One of the main solutions to the problem of reducing the dimension is to use feature selection. During the attribute selection process, a subset of the primary attributes is selected by removing the irrelevant and redundant attributes. The entire search space investigates to find the most appropriate feature subset.

In this paper, the initial dimension of a dataset is reduced using a graph-based feature selection method. Figure 2 shows the general flowchart of the proposed method.

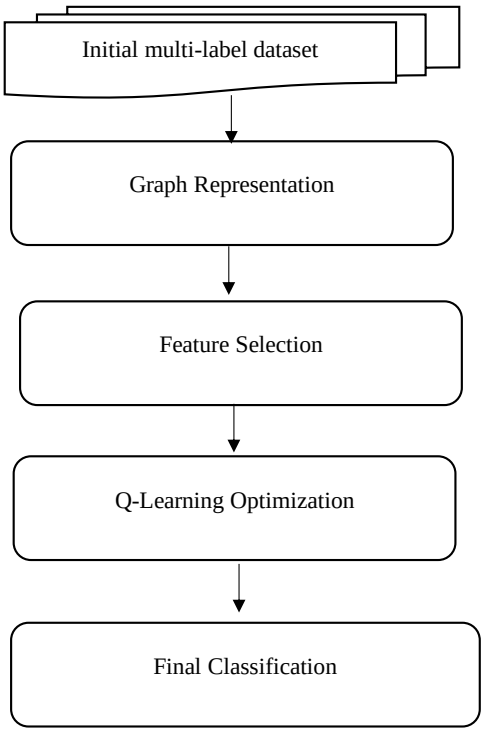

Figure 2: The flowchart of the proposed method.

To apply the graph-based method, the solution space of feature selection must be demonstrated by a weighted graph. To this end, in the first step, the initial features are shown with the Graph=$<G, E, w_{ij}>$, where $G=\{G_1, G_2, \ldots, G_n\}$ is a set of original features in which each feature shows a node in the graph, $E=\left\{(G_i, G_j) : G_i, G_j \in G\right\}$ denotes the set of edges of the graph, and w_{ij} indicates the similarity between two features G_i and G_j that are linked by the edge (G_i, G_j). This study employs the Pearson similarity criteria [15] to compute the similarity value of different features. The Pearson similarity

between two features G_i and G_j is shown by

$$W_{ij} = \left| \frac{\sum_p (x_i - \overline{x_i})\left(x_j - \overline{x_j}\right)}{\sqrt{\sum_p (x_i - \overline{x_i})^2 \ \sum_p \left(x_j - \overline{x_j}\right)^2}} \right|, \tag{1}$$

where x_i and x_j signify the feature vectors G_i and G_j, respectively. Variables $\overline{x_i}$ and $\overline{x_j}$ indicate the averages of vectors x_i and x_j over p samples. Greater similarity between the two features makes the Pearson criterion between the two features closer to 1, and reciprocally, the dissimilarity of the two features makes the Pearson criterion of the two features closer to 0.

Once the Pearson correlation coefficient is calculated, a normalization technique called SoftMax scaling [27] can be utilized to normalize these similarity values into the range from 0 to 1:

$$\widehat{w_{ij}} = \frac{1}{1 + exp(-\frac{w_{ij} - \overline{w}}{\sigma})}. \tag{2}$$

Here, w_{ij} is the similarity value between features G_i and G_j, $\overline{w}$ and σ are the average and variance of all of the Pearson correlations, respectively, and $\hat{w}_{ij}$ denotes the normalized correlation between features G_i and G_j.

In the second stage of the proposed feature selection, the weighted features are grouped in several clusters. The main purpose of feature clustering is to divide the primary features into a number of different clusters based on their similarity. Therefore, the features in each cluster are more similar to each other, and the features in different clusters are less similar to each other. In most feature clustering methods, the number of clusters must be determined before performing the clustering algorithm [21]. In other words, in most of these methods, the parameter k, which specifies the number of clusters, must be specified by the user. In general, it is difficult to determine the number of clusters for the initial characteristics of the work, and the number of optimal clusters can be determined only by trial and error.

For this reason, in this paper, a community detection algorithm called *Louvain* [3] is used for feature clustering. The present algorithm could detect the available communities existing in the graph through the maximization of a modularity function. This method is uncomplicated, effective, and easy-to-implement; one that could be utilized for the identification of communities in large networks. The complexity of computation for the algorithm is $O(n \log n)$, where n is the number of nodes in the graph. Therefore, it has the potential to be utilized in the detection of communities in extra-large networks in a short computing timespan. To maximize the specific network modularity, in the first step, each of the nodes is allocated to a selected community. In the second step, a new network is being made simply through the merging of previously detected communities. Then, until an imperative enhancement in the network modularity is achieved, the so-called process iterates. The present method has two main advantages. The first advantage is its intuitive and easy-to-implement stages, while the second is that it is exceedingly fast.

In the third stage, the appropriate features of each cluster are selected. The aim of this stage of the proposed method is to search for the optimal feature subset using

the concept of term variance (TV). In other words, in this step, an attempt is made to select each cluster of a number of features that are well able to represent all the features of that cluster. Clustering features and selecting the most effective features from each cluster ensure that the selected features provide a good representation of all the primary features. To select the most important features from each cluster, the TV criterion is utilized. This is a widely used criterion having low computational complexity and high efficiency. The TV criterion indicates the power of the attributes. Therefore, attributes having high scores provide valuable information. The TV criterion is defined by

$$TV(i) = \frac{1}{|G|} \sum_{j=1}^{|G|} \left(Attr(j,i) - \overline{Attr(i)} \right)^2, \tag{3}$$

where $Attr(j,i)$ indicates attribute j of sample i in the dataset, and $|G|$ denotes the number of all features in the dataset.

After the feature selection step, it is time for the final classification step. At this stage, the learning algorithm is used to deal with the problem of multi-label classification. The Q-learning algorithm is an extended version of the iterative value algorithm that is also used for uncertain problems. The Q-learning is a type of non-model reinforcement learning techniques that is based on dynamic random planning. In Q-learning, instead of defining a mapping from states to their values, a mapping is defined from the state/action pair to values called Q-values. Moreover, instead of defining a mapping from states to their values, a mapping is defined from the state/operation pair to the Q-values.

In this paper, we use Q-learning in multi-label classification problems by making changes to the algorithm structure. In the proposed algorithm, the search space is divided into several parts. Additionally, three new features are defined in the proposed algorithm.

In this paper, for the first time, using a combination of the TV feature selection algorithm and Q-learning, as a novel method, is proposed for feature selection. In the proposed method, unlike the existing methods, the final features in the problem of multi-label classification are optimally selected. The next steps will be decided by the agent, following Δt steps into the future. The weight for this step is calculated as $\gamma^{\Delta t}$, where γ is the discount factor whose value lies between 0 and 1 ($0 \leq \gamma \leq 1$), and demonstrates the valuing reward impacts which are received earlier and are more than the ones received later (reflecting a "good start" value). Additionally, γ might also be construed as the success (or survive) probability at each step Δt.

After selecting an appropriate subset of features, these features are used as the inputs of the classification model. It is worth mentioning that accurate feature selection and the use of an efficient prediction model to analyze the data lead to a better classification. A variety of machine learning algorithms, discussed in the previous section, have been proposed for classification. Machine learning classification algorithms such as support vector machine, artificial neural networks, deep learning, fuzzy systems, and ensemble learning models are used for multi-label classification.

The appropriate choice of a classifier is a vital step in classification. In the proposed method, a hybrid prediction model combining six different classifiers is used to take

advantage of all these classifiers. The hybrid model is an efficient prediction model that improves the total classification accuracy. In other words, the main idea of this paper is to use multiple classifiers in data classification. Consequently, the proposed hybrid model is very reliable when working with multi-label datasets.

In the proposed method, a combination of Q-learning, Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT) and K-Nearest Neighbors (KNN) classifiers is used in the classification process. Each of the classifiers makes its own prediction independently, and finally, based on the majority prediction, the final prediction is made. Given that each of the classifiers contributes to the final prediction, and according to the predicted emotional state by the majority, the final human state is recognized. This method is called Majority-Voting Classification (MVC). The pseudo-code of the proposed method is as follows.

Algorithm 1

Input D_T: Dataset
θ: Threshold for similarity
Output Predicted labels
Begin algorithm
$\boldsymbol{Graph}$ = Create a graph using Pearson similarity
$\boldsymbol{Graph}$ = Apply the threshold value θ to a primary graph
$\boldsymbol{G'}$= An empty set
Do
Calculate Feature Relevance
Sort features
Select the Best Feature $\boldsymbol{G'}$
Remove $\boldsymbol{G'}$ from the $\boldsymbol{Original\ Features}$
While (at least one feature has been found**)**
Report $\boldsymbol{G'}$ as a final feature set
Send the selected features to the Multi-Voting Classification
Specify the final classification using the Multi-Voting Classification
End algorithm

Figure 3: Pseudo-code of the proposed classification model

4 Experimental Results

To evaluate the performance of the proposed classification model on the multi-label classification problem, the proposed method was applied to various multi-label datasets. Furthermore, the obtained results were compared with those of the state-of-the-art works based on different metrics. The experiments were conducted in two parts.

MATLAB programming language was used to implement the feature selection method and other methods, and the results were derived based on these implementations. Also, all the tests were performed on a system with a 2.3 GHz Corei3 processor and 2 GB of internal memory (RAM).

In experiments on the proposed method, the dataset was randomly divided into training data and test data. For this purpose, 70% of the dataset was considered as the training data, and the remaining 30% as the test data. Also, in all the experiments, after identifying the train and test sets, each classification method was performed ten

times, and an average of the ten different executions was used to compare different methods.

4.1 The First Part of Experiments

In the first part, the algorithms were evaluated on the Emotions, Scene, Yeast and Genbase datasets according to the Sensitivity (%), Specificity (%), and Classification (%) metrics [32].

Another criterion used in this set of experiments was *average execution time*. In fact, we used this criterion to examine the complexity of the proposed method and the competitor methods. It is clear that less execution time indicates lower complexity and higher efficiency.

In Tables 1-4, the proposed method, which uses the graph-based feature selection method and Q-Learning, is compared with Multi-Label Classification with weighted classifier [30]. In these tables, Sensitivity, Specificity, and Classification rates are used to investigate the performance of the proposed method and the base methods existing in the literature.

Table 1: Comparison of the proposed method with multi-label classification with weighted classifier on the Emotions dataset

	Multi-label classification with weighted classifier			The proposed method		
	Best	Worst	Average	Best	Worst	Average
Sensitivity (%)	84.71	81.48	83.721	86.08	81.92	85.19
Specificity (%)	86.79	83.92	85.99	87.12	83.38	86.73
Classification (%)	85.51	81.31	84.42	86.39	82.54	85.62

Table 2: Comparison of the proposed method with multi-label classification with weighted classifier on the Scene dataset

	Multi-label classification with weighted classifier			The proposed method		
	Best	Worst	Average	Best	Worst	Average
Sensitivity (%)	82.62	81.54	83.71	86.89	80.38	85.54
Specificity (%)	87.78	83.29	84.28	87.99	85.82	84.71
Classification (%)	86.12	82.31	84.09	87.08	83.64	84.52

According to the results of Tables 1-4, the proposed method outperformed competitor algorithms and obtained better results.

Moreover, Figure 4 shows the average accuracy of the proposed method on different datasets. As shown in the figure, the proposed method has a higher average accuracy compared to the base methods existing in the literature.

Also, Table 5 allows us to compare the average execution time of the proposed method with that of multi-label classification with weighted classifier.

Table 3: Comparison of the proposed method with multi-label classification with weighted classifier on the Yeast dataset

	Multi-label classification with weighted classifier			The proposed method		
	Best	Worst	Average	Best	Worst	Average
Sensitivity (%)	84.61	81.41	83.22	86.28	82.91	84.12
Specificity (%)	86.89	82.04	84.21	87.46	83.87	84.23
Classification (%)	85.09	80.28	83.08	86.39	84.48	85.62

Table 4: Comparison of the proposed method with multi-label classification with weighted classifier on the Genbase dataset

	Multi-label classification with weighted classifier			The proposed method		
	Best	Worst	Average	Best	Worst	Average
Sensitivity (%)	84.72	80.42	83.71	86.87	81.91	85.98
Specificity (%)	85.38	82.35	84.72	88.59	84.87	86.71
Classification (%)	84.51	80.87	82.32	86.02	82.43	84.43

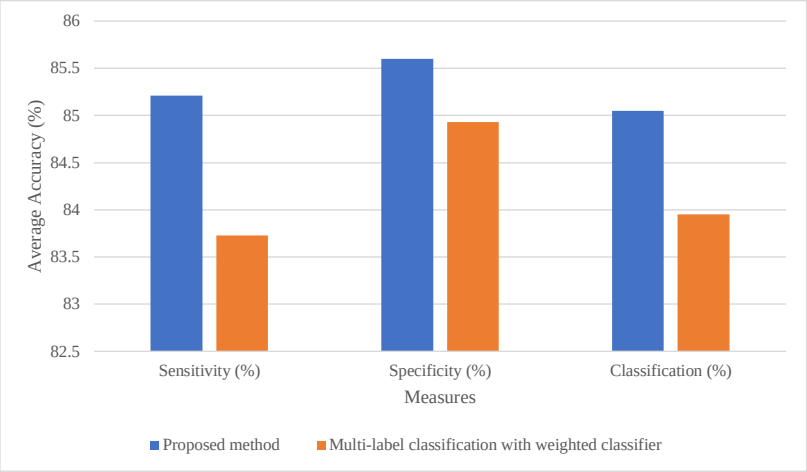

Figure 4: Comparison of the algorithms according to the average accuracy

The results of this experiment show that the proposed method has less execution time.

4.2 The Second Part of Experiments

In this subsection, the proposed multi-label classification is compared with the BR, MDDM, LIFT, LLSF, and MLSF methods in terms of Exact-Match, Hamming-Score, Macro-F1, and Micro-F1 metrics. The results of the competitor algorithms are adopted from the original paper which is reported by the authors, namely, [25]. It is worth

Table 5: Comparison of the expectation time for the proposed method and multi-label classification with weighted classifier

Execution Time	Multi-label classification with weighted classifier	The proposed method
1	119	89
2	118	82
3	112	80
4	115	85
5	118	91
6	125	79
7	113	83
8	108	84
9	118	92
10	121	81
Average	114.7	84.6

mentioning that all the details of the methods and datasets can be found in [25]. Tables 6-9 present the comparison results.

Table 6: Results obtained by the algorithms regarding the Exact-Math metric

Dataset	Methods					
	BR	MDDM	LIFT	LLSF	MLSF	Proposed
Emotions	0.285	0.263	0.184	0.285	0.315	**0.3305**
Scene	0.533	0.529	0.637	0.531	0.637	**0.7006**
Yeast	0.148	0.137	0.154	0.148	0.212	**0.2187**
Genbase	**0.982**	0.980	0.953	**0.982**	**0.982**	0.9474
Medical	0.665	0.609	0.574	0.662	0.689	**0.7013**
Enron	0.111	0.121	0.116	0.111	0.122	**0.1301**
Mediamill	0.066	0.068	0.069	0.066	0.070	**0.0715**
Bibtex	0.143	0.143	0.139	0.144	0.143	**0.1474**
Corel16k1	0.006	0.000	0.000	0.006	0.008	**0.0104**
Corel16k2	0.004	0.001	0.000	0.004	0.007	**0.0091**

According to Tables 6-9, the proposed multi-label classification model obtained better results compared to the competitor algorithms regarding the Exact-Match, Hamming-Score, Macro-F1, and Micro-F1 metrics.

A more detailed examination of the tables demonstrates that in the Genbase dataset, some competitor algorithms obtained better results compared to the proposed algorithm. Additionally, in the Enron and Mediamill datasets, the LIFT and MDDM methods obtained better values for the Hamming-Score and Micro-F1 metrics, respectively.

Table 7: Results obtained by the algorithms regarding the Hamming-Score metric

Dataset	Methods					
	BR	MDDM	LIFT	LLSF	MLSF	Proposed
Emotions	0.805	0.788	0.755	0.805	0.793	**0. 8136**
Scene	0.895	0.899	0.919	0.895	0.891	**0.9356**
Yeast	0.801	0.798	0.804	0.801	0.789	**0.8195**
Genbase	**0.999**	**0.999**	0.998	**0.999**	**0.999**	0.9964
Medical	0.990	0.988	0.987	0.990	0.990	**0.9931**
Enron	0.940	0.953	**0.955**	0.940	0.940	0.9504
Mediamill	0.968	0.969	0.969	0.968	0.968	**0.9706**
Bibtex	0.984	0.988	0.988	0.984	0.984	**0.9901**
Corel16k1	0.980	0.981	0.981	0.980	0.980	**0.9901**
Corel16k2	0.981	0.983	0.983	0.981	0.981	**0.9900**

Table 8: Results obtained by the algorithms regarding the Hamming-Score metric

Dataset	Methods					
	BR	MDDM	LIFT	LLSF	MLSF	Proposed
Emotions	0.633	0.583	0.496	0.633	**0.657**	0.6438
Scene	0.694	0.684	0.759	0.693	0.699	**0.8080**
Yeast	0.322	0.318	0.319	0.322	0.346	**0.3869**
Genbase	0.761	0.754	0.704	**0.769**	**0.769**	0.6080
Medical	0.366	0.323	0.240	0.370	0.387	**0.4094**
Enron	0.222	0.201	0.136	0.222	0.221	**0.2455**
Mediamill	0.028	0.035	0.035	0.028	0.029	**0.0364**
Bibtex	0.328	0.159	0.145	0.329	0.328	**0.3331**
Corel16k1	0.047	0.008	0.003	0.045	0.047	**0.0548**
Corel16k2	0.051	0.012	0.004	0.049	0.051	**0.0534**

Table 9: Results obtained by the algorithms regarding the Hamming-Score metric

Dataset	Methods					
	BR	MDDM	LIFT	LLSF	MLSF	Proposed
Emotions	0.661	0.627	0.557	0.661	0.665	**0.6749**
Scene	0.688	0.682	0.755	0.686	0.692	**0.8009**
Yeast	0.631	0.627	0.632	0.631	0.639	**0.6726**
Genbase	**0.993**	0.992	0.980	**0.993**	**0.993**	0.9664
Medical	0.810	0.780	0.679	0.804	0.815	**0.8331**
Enron	0.515	0.579	0.570	0.515	0.515	**0.5987**
Mediamill	0.510	**0.528**	0.519	0.510	0.491	0.5179
Bibtex	0.422	0.364	0.338	0.423	0.423	**0.4481**
Corel16k1	0.072	0.007	0.005	0.069	0.070	**0.0759**
Corel16k2	0.079	0.016	0.012	0.079	0.076	**0.0840**

5 Conclusion

Data mining refers to the study and analysis of large amounts of data in order to discover hidden and meaningful patterns and rules within them. Any dataset can be thought of as a valuable source of information, and the important point here is that the valuable information is hidden among a large amount of data, and we need to analyze the data to access that information. Classification is a form of data analysis that can be used to create a model to describe the data or to conceive directional mirror data. In multi-label classification, each data is associated with a subset of labels. This is called a set of related labels for that data. The aim of a multi-label learner is to create a function that maps each multi-label data to a set of related labels. Today, multi-label classifiers provide an important learning pattern among data mining learning algorithms. In this paper, a combination of graph-based feature selection and Q-learning, a new method was proposed to improve classification accuracy. The proposed algorithm was applied to ten well-known datasets, and the results were compared with those of the widely-used, state-of-the-art methods. The obtained results demonstrated that the proposed method had higher accuracy compared to the previous methods, and also had less computational complexity.

References

[1] Alizadeh J., Khaloozadeh H. (2019). "Enlarging the region of attraction for nonlinear systems through the sum-of-squares programming", Control and Optimization in Applied Mathematics, 4(2), 19-37.

[2] Al-Makhadmeh Z., Tolba A. (2019). "Utilizing IoT wearable medical device for heart disease prediction using higher order Boltzmann model: A classification approach", Measurement, 147, 106815.

[3] Blondel V., et al. (2008). "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment, 10008, 1-12.

[4] Boutell M.R., Luo J., Shen X., Brown C.M. (2004). "Learning multi-label scene classification", Pattern Recognition, 1757-1771.

[5] Dong H., et al. (2020). "A many-objective feature selection for multi-label classification", Knowledge-Based Systems, 208, 106456.

[6] 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.

[7] Hu J., et al. (2020). "Robust multi-label feature selection with dual-graph regularization", Knowledge-Based Systems, 203, 106126.

[8] Jayaraman V., Sultana H.P. (2019). "Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification", Journal of Ambient Intelligence and Humanized Computing.

[9] Ji Z., et al. (2020). "Deep ranking for image zero-shot multi-label classification", IEEE Transactions on Image Processing, 29, 6549-6560.

[10] Khammassi C., Krichen S. (2017). "A GA-LR wrapper approach for feature selection in network intrusion detection", Computers & Security, 70, 255-277.

[11] Khandagale S., Xiao H., Babbar R. (2020). "Bonsai: diverse and shallow trees for extreme multi-label classification", Machine Learning, 109(11), 2099-2119.

[12] Mafarja M., Mirjalili S. (2018). "Whale optimization approaches for wrapper feature selection", Applied Soft Computing, 62, 441-453.

[13] Makki I., Alhalabi W., Adham R.S. (2019). "Using emotion analysis to define human factors of virtual reality wearables", Procedia Computer Science, 163, 154-164.

[14] Mansourinasab S., Sojoodi M., Moghadasi S.R. (2019). "Model predictive control for a 3D pendulum on SO (3) manifold using convex optimization", Control and Optimization in Applied Mathematics, 4(2), 69-80.

[15] MonirulKabir Md., Shahjahan Md., Murase K. (2011). "A new local search based hybrid genetic algorithm for feature selection", Neurocomputing, 74(17), 2914-2928.

[16] Morais-Rodrigues F., et al. (2020). "Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression", Gene, 726, 144168.

[17] Naem A.A., Ghali N.I., Saleh A.A. (2018). "Antlion optimization and boosting classifier for spam email detection", Future Computing and Informatics Journal, 3(2), 436-442.

[18] Nasarian E., et al. (2020). "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach", Pattern Recognition Letters, 133, 33-40.

[19] Prajapati P., Thakkar A. (2021). "Performance improvement of extreme multi-label classification using K-way tree construction with parallel clustering algorithm", Journal of King Saud University - Computer and Information Sciences.

[20] Rejer I., Twardochleb M. (2018). "Gamers' involvement detection from EEG data with cGAAM – A method for feature selection for clustering", Expert Systems with Applications, 101, 196-204.

[21] Rostami M., Berahmand K., Forouzandeh S. (2020). "A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty", Journal of Big Data, 7(1), 83.

[22] Salman M.S., et al. (2019). "Group ICA for identifying biomarkers in schizophrenia: 'Adaptive' networks via spatially constrained ICA show more sensitivity to group differences than spatio-temporal regression", NeuroImage: Clinical, 22, 101747.

[23] Song D., et al. (2021). "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training", Information Systems, 101718.

[24] Song Q., et al. (2017). "Using deep learning for classification of lung nodules on computed tomography images", Journal of healthcare engineering, 2017.

[25] Sun L., Kudo M., Kimura, K. (2016, December). "Multi-label classification with meta-label-specific features", In 2016 23rd International Conference on Pattern Recognition (ICPR) (1612-1617). IEEE.

[26] Thabtah F.A., Cowling P., Peng Y., Rastogi R., Morik K., Bramer M., Wu X. (2004). "MMAC: A new multi-class, multi-label associative classification approach", Proc. Proceedings of Fourth IEEE International Conference on Data Mining, ICDM 2004, 217-224.

[27] Theodoridis S., Koutroumbas K. (2008). "Pattern recognition", Academic Press, Oxford.

[28] Wang W., et al. (2020). "MLCDForest: Multi-label classification with deep forest in disease prediction for long non-coding RNAs", Briefings in Bioinformatics.

[29] Wu T., et al. (2020). "Distribution-balanced loss for multi-label classification in long-tailed datasets", in European Conference on Computer Vision, Springer.

[30] Xia Y., Chen K., Yang Y. (2021). "Multi-label classification with weighted classifier selection and stacked ensemble", Information Sciences, 557, 421-442.

[31] Xu J. (2011). "An extended one-versus-rest support vector machine for multi-label classification", Neurocomputing, 3114-3124.

[32] Yang M., et al. (2019). "Investigating the transferring capability of capsule networks for text classification", Neural Networks, 118, 247-261.

[33] Yap X.H., Raymer M. (2021). "Multi-label classification and label dependence in in silico toxicity prediction", Toxicology in Vitro, 105157.

[34] You R., et al. (2020). "Cross-modality attention with semantic graph embedding for multi-label classification", in Proceedings of the AAAI Conference on Artificial Intelligence.

[35] Yun D., Ryu J., Lim J. (2021). "Dual aggregated feature pyramid network for multi label classification", Pattern Recognition Letters, 144, 75-81.

[36] Zheng X., et al. (2019). "Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning", Gene, 706, 188-200.

[37] Zhang Y., et al. (2020). "Large-scale multi-label classification using unknown streaming images", Pattern Recognition, 99, 107100.

[38] Zhong Y., Du B., Xu C. (2021). "Learning to reweight examples in multi-label classification", Neural Networks.

How to Cite this Article:
Seyed Ebrahimi, S.H., Majidzadeh, K., Soleimanian Gharehchopogh, F. (2022). "Multi-label classification with meta-label-specific features and Q-learning", Control and Optimization in Applied Mathematics, 6(2): 37-52. doi: 10.30473/coam.2021.60849.1177