Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 2814897, 12 pages
https://doi.org/10.1155/2018/2814897

Research Article

Feature Selection and Overlapping Clustering-Based Multilabel Classification Model

Liwen Peng and Yongguo Liu

Knowledge and Data Engineering Laboratory of Chinese Medicine, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Correspondence should be addressed to Yongguo Liu; liuyg_cn@163.com

Received 4 June 2017; Revised 2 November 2017; Accepted 10 December 2017; Published 4 January 2018

Academic Editor: Bogdan Dumitrescu

Multilabel classification (MLC) learning, which is widely applied in real-world applications, is a very important problem in machine learning. Some studies show that a clustering-based MLC framework performs effectively compared to a nonclustering framework. In this paper, we explore the clustering-based MLC problem. Multilabel feature selection also plays an important role in classification learning because many redundant and irrelevant features can degrade performance and a good feature selection algorithm can reduce computational complexity and improve classification accuracy. In this study, we consider feature dependence and feature interaction simultaneously, and we propose a multilabel feature selection algorithm as a preprocessing stage before MLC. Typically, existing cluster-based MLC frameworks employ a hard cluster method. In practice, the instances of multilabel datasets are distinguished in a single cluster by such frameworks; however, the overlapping nature of multilabel instances is such that, in real-life applications, instances may not belong to only a single class. Therefore, we propose a MLC model that combines feature selection with an overlapping clustering algorithm. Experimental results demonstrate that various clustering algorithms show different performance for MLC, and the proposed overlapping clustering-based MLC model may be more suitable.

1. Introduction

The multilabel classification (MLC) problem, which is applicable to a wide variety of domains, such as music classification and bioinformatics [1], has received increasing attention. However, situations where single instances are associated with multiple labels remain challenging. Most algorithms treat such MLC tasks as multiple binary classification tasks. However, this approach may not consider potential correlations among features and labels.

A good MLC solution must be effective and efficient; however, a large number of redundant and irrelevant attributes may increase computational costs and the time required to learn and test a multilabel classifier, which reduces classification performance. Feature selection, which is an important technique in data mining and machine learning, has been widely used in classification models to enhance performance. Selecting features before applying classification methods to original datasets has many advantages, such as refining the data, reducing computational costs, and improving classification accuracy [2, 3]. Therefore, we utilise a feature selection algorithm to improve the quality of MLC.

Various feature selection methods have been proposed, for example, statistics, rough set methods, information gain, and mutual information (MI). A wide variety of research has shown that no single feature selection method can handle all situations. Many studies have demonstrated that MI-based feature selection methods are effective and efficient because the MI can handle different types of attributes, does not make any assumptions, and can measure nonlinear relations between variables [4]. Recently, many algorithms to select significant features for MLC have been proposed. However, most of these methods do not consider that a single attribute may affect various labels differently. The concept of interaction information has become more relevant because it can reflect the relevance, redundancy, and complementarity among attributes and labels; thus, it is an effective feature selection method. In this study, we propose an algorithm to

improve MLC performance by selecting significant attributes based on interaction information between attributes and labels.

Some studies have shown that clustering-based MLC methods can improve predictive performance and reduce time costs; however, those studies used nonoverlapping clustering methods to handle multilabel datasets. We know that, in MLC, one object may belong to multiple classes; however, algorithms based on nonoverlapping clustering, that is, hard division methods, do not consider such situations. In contrast, overlapping clustering-based methods consider this situation when they handle datasets. Therefore, we propose an overlapping clustering-based MLC (OCBMLC) model.

The remainder of this paper is organised as follows: Section 2 describes related work, Section 3 provides background information, Section 4 describes the proposed multilabel feature selection algorithm and MLC model, Section 5 introduces experimental data, evaluation criteria, and experimental results, and conclusions and suggestions for future work are presented in Section 6.

2. Related Work

Currently, a variety of algorithms have been developed to handle MLC problems [5–15]. In traditional classification methods, each instance has a single label; however, in MLC, an instance can have more than one label. MLC algorithms can be divided into problem transformation methods (PT) and algorithm adaptation methods (AA) [9].

PT methods convert multilabel data to single-label data; thus, a single-label classification method can be used. Label powerset, binary relevance [10], and random ensemble learning with k-label sets [11] are classic PT methods. The AA approaches extend single-label algorithms to process multilabel data directly. BP-MLL [12] and ML-KNN are two popular AA methods. BP-MLL is a widely used MLC backpropagation algorithm. An important characteristic of this algorithm is the introduction of an error function that considers multiple labels. The ML-KNN AA method [13] determines the labels of a new object using the maximum a posteriori principle. The ML-KNN algorithm obtains a label set based on the statistical information of the label sets of the k-nearest neighbours of a test instance.

Many studies have proven that redundant and irrelevant features can increase computational costs, reduce performance, and result in overfitting. These problems also exist in MLC. Many feature selection methods have been proposed to handle these problems and improve MLC. Battiti [14] proposed the Mutual Information Feature Selection algorithm, which selects the maximum relevance term, to address these problems. Peng et al. [15] introduced an improvement algorithm called Minimal-Redundancy and Max-Relevance, and Lin et al. [16] proposed a multilabel feature selection algorithm that combines MI with max-dependency and min-redundancy. In addition, over the past few years, unsupervised, clustering, and other technologies have been used to reduce dimensionality. For example, Li et al. [17] proposed a clustering-guided sparse structural learning algorithm that integrates clustering and a sparse structure in a united framework to select the most useful features. They also proposed an algorithm [18] that employs nonnegative spectral clustering and controls the redundancy between features to select significant features. Cai et al. [19] presented the Unified Sparse Subspace Learning (USSL) framework, which employs a dimension reduction technique that incorporates a subspace learning method. The USSL framework has demonstrated good performance. Li et al. [20] proposed the Robust Structured Subspace Learning (RSSL) framework that combines subspace learning theory and features learning. Their experimental results demonstrated that the RSSL framework performed well for image understanding tasks.

Recently, Kommu et al. [21] proposed two methods based on probabilistic theory to solve multilabel learning problems. In the first method, their algorithm uses logistic regression and a nearest neighbour classifier for MLC. Note that Partial Information is used in this approach. In the second method, their algorithm deals with the concept of grouping related labels. Association Rules are also introduced in the second approach. Guo and Li [22] proposed the Improved Conditional Dependency Networks framework for MLC. This method uses label correlations in the training stage and CDNs in the testing stage. Yu et al. [23] used a rough sets approach for MLC that considers the associations between labels. They evaluated the performance of their approach using seven multilabel datasets.

Nasierding et al. [24, 25] presented an effective CBMLC framework that combines a clustering algorithm with an MLC algorithm. Various clustering methods, such as k-means, EM, and Sequential Information Bottleneck, are used for training. Note that, with this framework, labels are ignored during training phase. Nasierding et al. [26] compared clustering and nonclustering MLC methods for image and video annotations. Tahir et al. [27] proposed a method that combines a multilabel learning approach with fusion techniques. They used various multilabel learners to select a label set and demonstrated that ensemble techniques can avoid the disadvantages of different learners.

3. Background Theory

3.1. Entropy and Mutual Information. In this section, we introduce the theories of Entropy and MI. Here, we assume that all variables are discrete or data attributes can be discretised using different discrete methods. Shannon's entropy [28] is the uncertainty measure of a random variable, and it has been widely used in various domains. Here, let $X \in \{x_1, x_2, \ldots, x_n\}$ be a discrete variable and $p(x_i) = \text{probability}\{X = x\}$ be the probability density function. Formally, the entropy of X is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log\left(p(x_i)\right). \tag{1}$$

Assume that X and $Y \in \{y_1, y_2, \ldots, y_m\}$ are two random discrete variables. $p(x_i, y_j)$ is the joint probability of X and Y. The joint entropy $H(X, Y)$ is defined as follows:

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p\left(x_i, y_j\right) \log\left(p\left(x_i, y_j\right)\right). \tag{2}$$

If the value of random variable X is known and variable Y is not, the remaining uncertainty of variable Y can be measured by the conditional entropy defined as follows:

$$H(Y \mid X) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p\left(y_j, x_i\right) \log\left(p\left(y_j \mid x_i\right)\right). \quad (3)$$

The minimum value of $H(Y \mid X)$ is zero when random variable Y is statistically dependent on random variable X. The maximum conditional entropy value occurs when the two variables are statistically independent.

The relationship between conditional and joint entropy can be defined as follows:

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y). \quad (4)$$

MI is the amount of information shared by two variables and is defined as follows:

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i, y_j\right) \log\left(\frac{p\left(x_i, y_j\right)}{p\left(x_i\right) P\left(y_j\right)}\right). \quad (5)$$

Note that the two random variables are statistically independent when $I(X,Y)$ is zero. The relation between MI and entropy can be defined as follows:

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \quad (6)$$

Let Z be a random variable $Z = \{z_1, z_2, \ldots, z_k\}$. The conditional MI and joint MI represent the information of two variables in the context of a third variable and are defined as follows:

$$\begin{aligned} I(X; Z \mid Y) &= H(X \mid Z) - H(X \mid Z, Y) \\ I(X, Y; Z) &= I(X; Z \mid Y) - H(Y, Z). \end{aligned} \quad (7)$$

Multi-information, which was introduced by McGill [29], is an extension of two random variables that can handle the interaction among more than two random variables. Mathematically, multi-information is defined as follows:

$$I(X; Y; Z) = I(X, Z; Y) - I(X, Z) - I(Y; Z). \quad (8)$$

Multi-information can be positive, negative, or zero [30]. If the multi-information value is zero, the random variables are independent in the context of the third variable. If the value is negative, the variables have redundant information and a positive value indicates that together the random variables can provide more information than each variable taken individually.

3.2. Overlapping Clustering Algorithm. Fuzzy C-Means (FCM) algorithms are widely used in fuzzy clustering learning. Fuzzy clustering, which is a type of overlapping clustering, differs from hard clustering. The FCM clustering algorithm assigns data points (examples) to a cluster, and the fuzzy membership of data points indicates the extent to which data points pertain to their clusters [31].

Suppose $X \in \{x_1, x_2, \ldots, x_n\}$ is a set of n vectors for c clustering. Vectors $x_i \in R^s$ represent the attributes of the object x_i. Here, a fuzzy partition matrix M_{fc} $(c \times n)$ is defined as

$$\begin{aligned} M_{fc} = \Bigg\{ & W \in R^{cn} \mid w_{ik} \in [0,1],\ \forall i,k;\ \sum_{i=1}^{c} w_{ik} \\ & = 1,\ \forall k;\ 0 < \sum_{k=1}^{n} w_{ik} < n,\ \forall i \Bigg\}, \end{aligned} \quad (9)$$

where $1 \le i \le c,\ 1 \le k \le n$.

Note that examples can belong to more than one cluster with different degrees of membership. The object function of the FCM algorithm obtains the minimum value as follows [32]:

$$\min J_m(W,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} w_{ij}^{m} d_{ij}^{2}\left(x_i, v_i\right), \quad (10)$$

where $W = \{w_{ij}\}$ is the membership degree matrix, parameter $m \in (1, +\infty)$ is the weight exponent that defines the fuzziness of the resulting clusters, and $d_{ij} = \|x_k - p_i\|$ is the Euclidian distance between object x_k and the cluster centre v_i.

The objective function is minimized by updating the partition matrix and cluster centre as follows:

$$\begin{aligned} v_i &= \frac{\sum_{k=1}^{n} \left(w_{ik}\right)^m x_k}{\sum_{k=1}^{n} \left(w_{ik}\right)^m} \\ {w_{ik}}^{b} &= \sum_{j=1}^{c} \left[\left(\frac{d_{ik}^{(b)}}{d_{jk}^{(b)}} \right)^{2/(m-1)} \right]^{-1}. \end{aligned} \quad (11)$$

The FCM membership function is defined as follows:

$$u_{i,j} = \sum_{t=1}^{c} \left[\left(\frac{\left\| x_j - v_i \right\|_A}{\left\| x_j - v_t \right\|_A} \right)^{2/(m-1)} \right]^{-1}. \quad (12)$$

Here, $u_{i,j}$ is the membership value of the jth object and ith cluster, c is the number of clusters, and v_i is the cluster centre of the ith cluster.

4. Proposed Multilabel Classification Model

4.1. Proposed Multilabel Feature Selection Method. In information theory, $F = \{f_1, f_2, \ldots, f_n\}$ is the original feature set, and the subset $S = \{f_1, f_2, \ldots, f_k\}$ is the compact feature subset where $n > k$. The selected subset $S = \{f_1, f_2, \ldots, f_k\}$ should maximise the joint information between the subset S with compact dimension k and the class label C.

$$\mathrm{MI}(F; C) = \mathrm{MI}\left(f_1, f_2, \ldots, f_k; C\right). \quad (13)$$

Such a method is impractical because it is difficult to calculate the high-dimensional probability density function. Therefore, some efficient methods have been proposed to approximate the ideal solution [14–16]. Generally, most multilabel feature selection methods based on MI consider the

relevance and redundancy terms. In practice, such methods and their variants calculate the MI between a candidate feature and the selected features subset; however, they do not sufficiently consider interaction information among attributes and class labels, ignore feature cooperation, and allow all features to be competitive.

We know that a candidate feature for multilabel feature selection should have one of the highest MI values for all class labels. This is referred to as the relevance term. Multilabel feature relevance terms have been defined previously, and we use the following definition.

Definition 1. Let f_i denote a candidate feature and $l_k \in L$ be a class label. The relevance term is expressed as follows:

$$D = \sum_{l_k \in L} \mathrm{MI}\left(f_i; l_k\right). \tag{14}$$

We can obtain two properties according to this definition.

Property 2. If candidate feature f_i and each class label $l_k \in L$ are mutually independent, then the MI of f_i and L is minimum.

Property 3. If each class label $l_k \in L$ is determined completely by f_i, then the MI of f_i and L is maximum.

According to the above properties, we can use Definition 1 to select relevant candidate features. However, classes combined with previously selected features may produce interaction. Therefore, we should consider the interaction information among the candidate feature, the selected features, and the classes during feature selection. Differing from existing feature selection methods, we consider the interaction information between a feature and a single class and the pairwise interaction between features and all class labels. Our interaction metric is defined as follows:

$$I = \sum_{l_k \in L} \sum_{f_j \in S} \mathrm{MI}\left(f_i; f_j; l_k\right). \tag{15}$$

Here, S is the selected features subset, L denotes the label set, and f_i denotes the candidate feature.

It is well known that multilabel feature selection attempts to select a set of features with the highest discrimination power for all labels. According to the above discussion, we combine (14) and (15) using the feature interaction maximum of the minimum criteria to propose a new goal function (referred to as max-dependence and interaction (MDI)) for multilabel feature selection. Here, the candidate features are considered to have the highest relevance and beneficial interaction with all class labels. The proposed MDI goal function is expressed as follows:

$$\mathrm{MDI} = \arg\max \left[\sum_{l_k \in L} \mathrm{MI}\left(f_i; l_k\right) + \min \sum_{l_k \in L} \sum_{f_j \in S} \mathrm{MI}\left(f_i; f_j; l_k\right) \right]. \tag{16}$$

With this function, the first term is the relevance between the candidate features and all class labels, and the second term focuses on the interaction information among f_i, f_j, and L. The proposed goal function can select features with the greatest discrimination power. The pseudocode of the proposed algorithm is as Pseudocode 1.

4.2. Proposed Multilabel Classification Model. There are some experimental results that show that CBMLC methods can improve the predictive classification performance and reduce algorithm training time compared to existing popular multilabel methods [24–26]. The results of those models show that the classification performance of clustering-based methods is effective. However, those algorithms were used for nonoverlapping clustering methods, such as EM and k-means, prior to MLC. Therefore, the original data will be set into several disjoint data clusters in nonoverlapping methods.

Clustering methods are usually classified into hard clustering and fuzzy clustering. In hard clustering, instances are distinguished in a single cluster. However, due to the overlapping nature of instances, generally, they do not belong to only a single class in real-world applications. This property limits the practical application of hard clustering, especially for MLC.

FCM is an effective classic fuzzy clustering method based on an objective function concept and is widely used in clustering. The FCM approach uses alternating optimisation strategies to solve nonlinear and nonmonitor clustering problems. We know that one instance may own multiple classes in multilabel data, and the FCM algorithm can handle one instance that belongs to more than one cluster simultaneously. This allows the use of a fuzzy clustering method that assigns a single object to several clusters. Therefore, we propose the OCBMLC model in combination with the FCM algorithm to improve performance. Figure 1 shows the basic procedure of the proposed OCBMLC model.

5. Experiments and Results

5.1. Datasets. In our experiments, we used three public multilabel datasets, that is, the emotions, yeast, and scene datasets. These datasets were taken from the Mulan Library. The emotions dataset contains examples of songs according to people's emotions [33]. The yeast dataset includes information about genes functions [34], and the scene [35] dataset includes a series of landscape patterns. Table 1 shows the statistics of the three multilabel benchmark datasets.

In Table 1, "Domain" denotes the dataset domain, "Instances" is the number of instances in the dataset, "Features" is the number of attributes, "Labels" is the number of labels in the datasets, and "Cardinality" is the average number of labels associated with each instance.

5.2. Experimental Setting. At the multilabel feature selection stage, in order to calculate MI convenience, we discretise continuous features into 10 bins using an equal-width strategy. The evaluation approaches for MLC differ from traditional single-label classifications. Note that the Hamming loss and

(1) (Initialisation) Set $X \leftarrow$ "initial set of n features";
$S \leftarrow$ "empty set".
(2) (Computation of the MI with the output class set)
For $\forall x_i \in X,\ \forall l_i \in L$ compute $\sum_{l_k \in L} \mathrm{MI}(x_i; l_k)$
(3) (Choice of the first feature) Find the feature x that
maximizes $\sum_{l_k \in L} \mathrm{MI}(x_i; L)$; set $X \leftarrow X \setminus \{x_i\}$; set $S \leftarrow \{x_i\}$.
(4) (Greedy selection) Repeat until $|S| = k$;
(selection of the next feature) choose the feature
$x_i = \arg\max[\sum_{l_k \in L} \mathrm{MI}(x_i; l_k) + \min \sum_{l_k \in L} \sum_{f_j \in S} \mathrm{MI}(x_i; x_j; l_k)]$;
set $X \leftarrow X \setminus \{x_i\}$; set $S \leftarrow S \cup \{x_i\}$;
(5) (output) Output the set S with the selected features.

Pseudocode 1

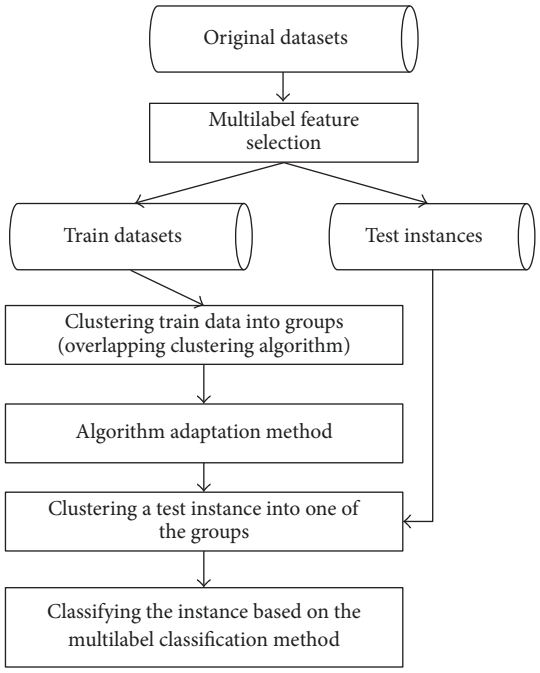

Figure 1: Basic procedure of OCBMLC model.

Table 1: The statistic of the multilabel benchmark datasets.

Dataset	Domain	Instances	Features	labels	Cardinality
Emotions	Music	593	72	6	1.869
Yeast	Biology	2417	103	14	4.237
Scene	Image	2407	294	6	1.074

micro $F1$-measure evaluation criteria are widely used for MLC; thus, we used these criteria in our experiments.

Note that non-OCBMLC models use k-means and EM algorithms to cluster original datasets, and OCBMLC model uses the FCM algorithm on the data after dimension reduction. The overlapping and nonoverlapping frameworks both employ ML-KNN as the classifier. The number of clusters k in k-means, EM, and FCM is all set between 2 to 7. In this study, a cross-validation strategy was used for each combination of algorithm framework and dataset. All experiments used MATLAB 2012 on an Intel Core-i5 2.3 GHz processor with 8 GB memory.

5.3. Evaluation Metrics. The evaluations of an MLC system differ from that of a single-label classification system. Note that some criteria that evaluate the performance of an MLC system have been employed previously [36]. Among such evaluation metrics, we employed Hamming loss and the micro $F1$-measure criteria.

Here, let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a set of n test examples and $Y^* = h(x_i)$ be the predict label set for the test instance X_i. Y_i is the ground truth label set for X_i. The Hamming loss indicates the number of erroneous labels to the total number of labels, where a smaller Hamming loss value indicates better classification performance. The Hamming loss value is calculated as follows:

$$\text{Hamming Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{Q} \cdot \sum_{q=1}^{Q}\left(\delta\left(q \in Y_i^* \wedge q \notin Y_i\right) + \delta\left(q \notin Y_i^* \wedge q \in Y_i\right)\right). \tag{17}$$

The micro $F1$-measure represents the harmonic means between precision and recall, and it is calculated from false positives, false negatives, true positives, and true negatives. The $F1$-measure and microaveraging are evaluated as follows:

$$F1\text{-measure} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\left|Y_i \cap Z_i\right|}{\left|Z_i\right| + \left|Y_i\right|}$$
$$M_{\text{micro}} = M\left(\sum_{l=1}^{M} tp_l, \sum_{l=1}^{M} fp_l, \sum_{l=1}^{M} tn_l, \sum_{l=1}^{M} fn_l\right). \tag{18}$$

Here, tp_l denotes true positives and fp_l denotes false positives, and tn_l and fn_l are true and false negatives, respectively, for l labels after a separate binary evaluation is performed. Note that a greater micro $F1$-measure value indicates better classification performance of a multilabel algorithm.

5.4. Results. In this study, we used Hamming loss and the micro $F1$-measure as experimental evaluation metrics and

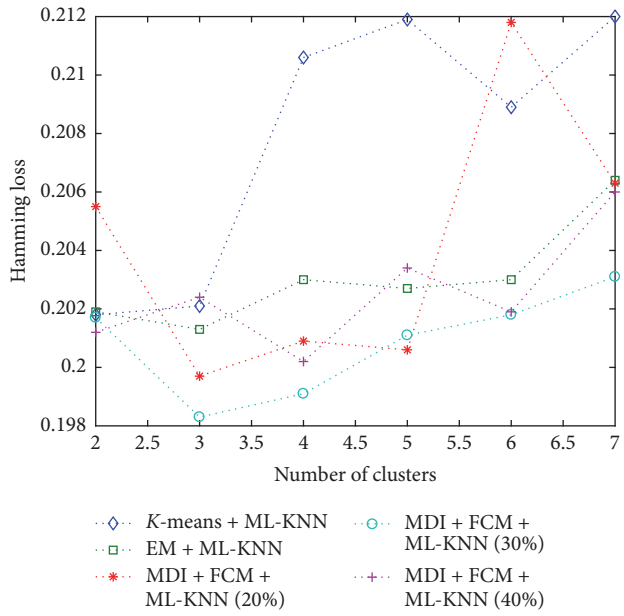

Figure 2: Hamming loss for all models and the number of clusters in emotions.

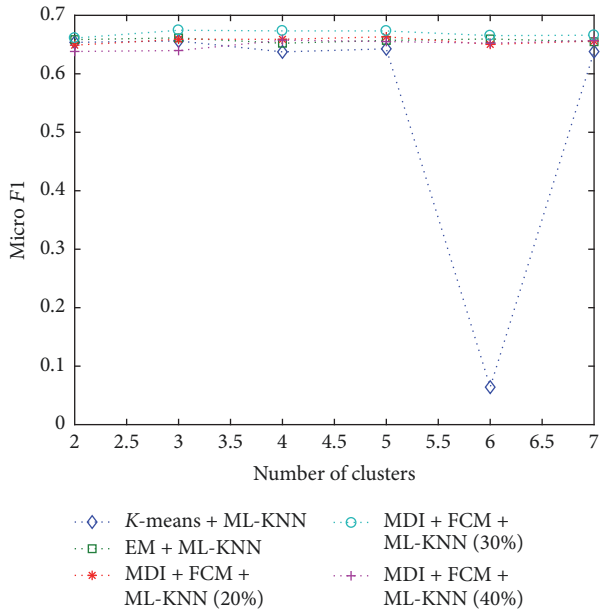

Figure 3: Micro $F1$ for all models and the number of clusters in emotions.

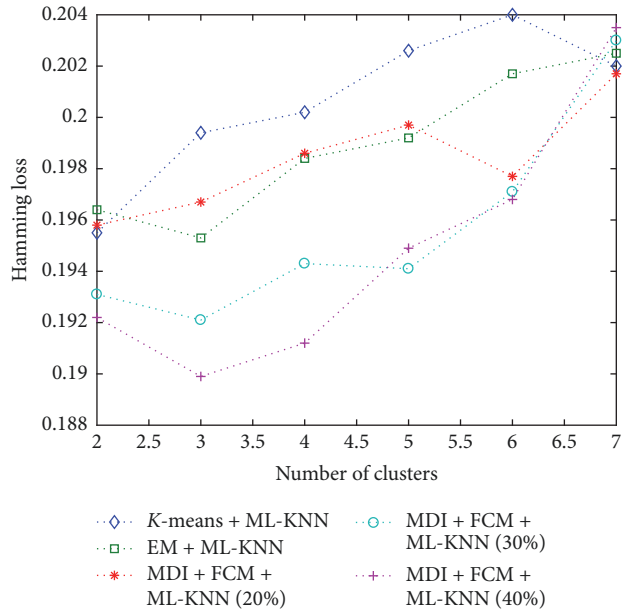

Figure 4: Hamming loss for all models and the number of clusters in yeast.

employed ML-KNN as the multilabel classifier. Note that, in all cases, we indicate the best results in bold values in Tables 2, 3, 4, 5, 6, 7, 8, and 9.

5.4.1. Comparisons of Feature Selection Methods. To demonstrate the efficacy of the proposed feature selection algorithm, we compared the proposed feature selection method to other MLC models based on clustering using the emotion dataset. We also compared a feature selection method that only considers the dependence between features and classes using the proposed algorithm in which interaction information among features and classes is considered. Here, we refer to the criterion that considers only dependence as the max-dependence criterion, where DEP_max $= \max \sum_{l_k \in L} \mathrm{MI}(f_i; l_k)$. This criterion was used to select candidate features.

In this experiment, "DEP_max" represents the features selected by the max-dependence criterion, which ignores interaction information when selecting candidate features, and "MDI" represents features selected by the proposed algorithm, which considers dependence and interaction

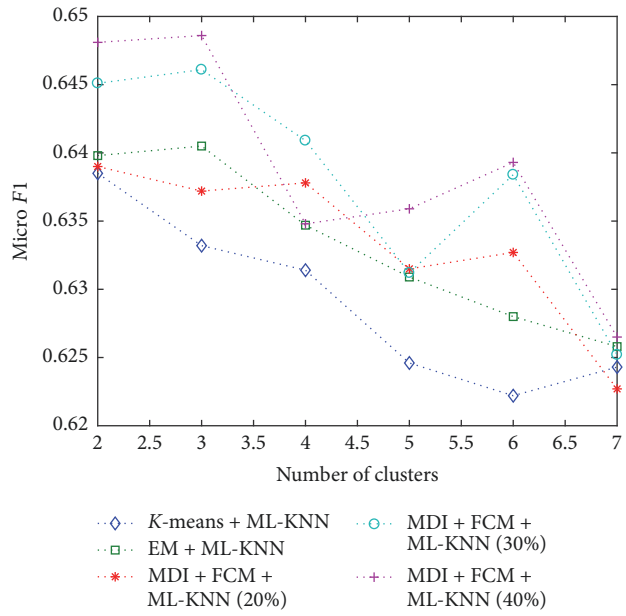

Figure 5: Micro $F1$ measure for all models and the number of clusters in yeast.

Table 2: Hamming loss measure for the models based on k-means clustering on emotions dataset. n is the number of attributes, and k is the amount of clusters. s% is the average value of Hamming loss.

Classification model	k	n	Emotions
k-means + ML-KNN	2	72	0.2018
EM + ML-KNN	2	72	0.2019
DEP_max + k-means + ML-KNN	2	s%	0.2021
MDI + k-means + ML-KNN	2	s%	0.2016
k-means + ML-KNN	3	72	0.2021
EM + ML-KNN	3	72	0.2013
DEP_max + k-means + ML-KNN	3	s%	0.2034
MDI + k-means + ML-KNN	3	s%	**0.1986**
k-means + ML-KNN	4	72	0.2106
EM + ML-KNN	4	72	0.2030
DEP_max + k-means + ML-KNN	4	s%	0.2156
MDI + k-means + ML-KNN	4	s%	0.2024
k-means + ML-KNN	5	72	0.2119
EM + ML-KNN	5	72	0.2027
DEP_max + k-means + ML-KNN	5	s%	0.2168
MDI + k-means + ML-KNN	5	s%	0.2005
k-means + ML-KNN	6	72	0.2089
EM + ML-KNN	6	72	0.2030
DEP_max + k-means + ML-KNN	6	s%	0.2078
MDI + k-means + ML-KNN	6	s%	0.2056
k-means + ML-KNN	7	72	0.2120
EM + ML-KNN	7	72	0.2064
DEP_max + k-means + ML-KNN	7	s%	0.2126
MDI + k-means + ML-KNN	7	s%	0.2030

Table 3: Micro $F1$-measure for all models on emotions dataset. n is the number of attributes, and k is the amount of clusters. s% is the average value of micro $F1$-measure.

Classification model	k	n	Emotions
k-means + ML-KNN	2	72	0.6538
EM + ML-KNN	2	72	0.6585
DEP_max + k-means + ML-KNN	2	s%	0.6572
MDI + k-means + ML-KNN	2	s%	0.6590
k-means + ML-KNN	3	72	0.6562
EM + ML-KNN	3	72	0.6614
DEP_max + k-means + ML-KNN	3	s%	0.6532
MDI + k-means + ML-KNN	3	s%	**0.6650**
k-means + ML-KNN	4	72	0.6375
EM + ML-KNN	4	72	0.6523
DEP_max + k-means + ML-KNN	4	s%	0.6350
MDI + k-means + ML-KNN	4	s%	0.6635
k-means + ML-KNN	5	72	0.6427
EM + ML-KNN	5	72	0.6571
DEP_max + k-means + ML-KNN	5	s%	0.6430
MDI + k-means + ML-KNN	5	s%	0.6590
k-means + ML-KNN	6	72	0.6411
EM + ML-KNN	6	72	0.6596
DEP_max + k-means + ML-KNN	6	s%	0.6423
MDI + k-means + ML-KNN	6	s%	0.6621
k-means + ML-KNN	7	72	0.6380
EM + ML-KNN	7	72	0.6549
DEP_max + k-means + ML-KNN	7	s%	0.6382
MDI + k-means + ML-KNN	7	s%	0.6576

information among the candidate features and each class simultaneously. Here, we selected the top n% ($n = 20, 30, 40$) features for MLC according to the MDI and DEP_max criteria, and we used the average Hamming loss and micro

Table 4: Hamming loss measure for all models on emotions dataset. n is the number of attributes, and k is the amount of clusters.

Classification model	k	n	Emotions
k-means + ML-KNN	2	72	0.2018 ± 0.0167
EM + ML-KNN	2	72	0.2019 ± 0.0265
MDI + FCM + ML-KNN	2	20%	0.2055 ± 0.0033
		30%	0.2017 ± 0.0151
		40%	0.2012 ± 0.0072
k-means + ML-KNN	3	72	0.2021 ± 0.0182
EM + ML-KNN	3	72	0.2013 ± 0.0204
MDI + FCM + ML-KNN	3	20%	0.1997 ± 0.0009
		30%	**0.1983 ± 0.0154**
		40%	0.2024 ± 0.0012
k-means + ML-KNN	4	72	0.2106 ± 0.0251
EM + ML-KNN	4	72	0.2030 ± 0.0261
MDI + FCM + ML-KNN	4	20%	0.2009 ± 0.0018
		30%	0.1991 ± 0.0108
		40%	0.2002 ± 0.0111
k-means + ML-KNN	5	72	0.2119 ± 0.0122
EM + ML-KNN	5	72	0.2027 ± 0.0289
MDI + FCM + ML-KNN	5	20%	0.2006 ± 0.0037
		30%	0.2011 ± 0.0075
		40%	0.2034 ± 0.0100
k-means + ML-KNN	6	72	0.2089 ± 0.0177
EM + ML-KNN	6	72	0.2030 ± 0.0239
MDI + FCM + ML-KNN	6	20%	0.2118 ± 0.0150
		30%	0.2018 ± 0.0054
		40%	0.2019 ± 0.0129
k-means + ML-KNN	7	72	0.2120 ± 0.0183
EM + ML-KNN	7	72	0.2064 ± 0.0243
MDI + FCM + ML-KNN	7	20%	0.2063 ± 0.0139
		30%	0.2031 ± 0.0041
		40%	0.2060 ± 0.0130

Table 5: Micro $F1$-measure for all models on emotions dataset. n is the number of attributes, and k is the amount of clusters.

Classification model	k	n	Emotions
k-means + ML-KNN	2	72	0.6538 ± 0.0211
EM + ML-KNN	2	72	0.6585 ± 0.0425
MDI + FCM + ML-KNN	2	20%	0.6490 ± 0.0295
		30%	0.6611 ± 0.0291
		40%	0.6382 ± 0.0286
k-means + ML-KNN	3	72	0.6562 ± 0.0340
EM + ML-KNN	3	72	0.6614 ± 0.0343
MDI + FCM + ML-KNN	3	20%	0.6591 ± 0.0419
		30%	**0.6745 ± 0.0265**
		40%	0.6398 ± 0.0181
k-means + ML-KNN	4	72	0.6375 ± 0.0432
EM + ML-KNN	4	72	0.6523 ± 0.0437
MDI + FCM + ML-KNN	4	20%	0.6590 ± 0.0420
		30%	0.6734 ± 0.0341
		40%	0.6575 ± 0.0165
k-means + ML-KNN	5	72	0.6427 ± 0.0240
EM + ML-KNN	5	72	0.6571 ± 0.0481
MDI + FCM + ML-KNN	5	20%	0.6633 ± 0.0170
		30%	0.6733 ± 0.0243
		40%	0.6552 ± 0.0237
k-means + ML-KNN	6	72	0.6411 ± 0.0227
EM + ML-KNN	6	72	0.6596 ± 0.0321
MDI + FCM + ML-KNN	6	20%	0.6303 ± 0.0458
		30%	0.6653 ± 0.0116
		40%	0.6524 ± 0.0155
k-means + ML-KNN	7	72	0.6380 ± 0.0222
EM + ML-KNN	7	72	0.6549 ± 0.0404
MDI + FCM + ML-KNN	7	20%	0.6561 ± 0.0248
		30%	0.6660 ± 0.0219
		40%	0.6568 ± 0.0233

$F1$-measure values based on the selected top n% (n = 20, 30, 40) features subset by comparing the values from the original feature sets.

Table 2 shows the Hamming loss values obtained when we used the features selected according to MDI, DEP_max, and the original features from the emotion dataset, and Table 3 shows the micro $F1$-measure values when we employed features selected according to MDI, DEP_max, and the original features. In terms of the feature selection methods, we found that the performance of DEP_max is no better than that of the other models even though we used the original feature subset. However, the MDI performance is better and more stable when the clustering number k is 2 to 7. It is likely that features selected by only considering Max-Relevance could generate abundant redundancy, which means that the dependence among those features could be large. Therefore, the proposed feature selection function may be better suited for MLC, and we observed the same from the experimental results.

5.4.2. Comparisons of Multilabel Classification Models. The results obtained by the MLC models with the emotions dataset relative to Hamming loss and the micro $F1$-measure are shown in Tables 4 and 5. We selected the top n% (n = 20, 30, 40) in the selected feature subset as the final feature subset for use with the proposed model. Table 4 demonstrates that the proposed OCBMLC framework achieved the lowest Hamming loss value (0.1983 ± 0.0154) with the emotions dataset. Table 5 shows that the proposed framework achieved the highest micro $F1$-measure (0.6745 ± 0.0265) with the emotions dataset. As shown in Figures 2 and 3, the predictive performance of the proposed model achieved the best results with the emotions dataset when k = 3. As shown in Figures 2 and 3, respectively, the Hamming loss demonstrates the minimum value and the micro $F1$-measure demonstrates the maximum value when we used the MDI feature selection method to select the top n% (n = 30) features as the classification attributes subset.

To demonstrate the classification performance of the proposed model, we also selected the top p% (p = 20, 30, 40)

Table 6: Hamming loss measure for all models on yeast dataset. n is the number of attributes, and k is the amount of clusters.

Classification model	k	n	Yeast
k-means + ML-KNN	2	103	0.1955 ± 0.0117
EM + ML-KNN	2	103	0.1964 ± 0.0117
MDI + FCM + ML-KNN	2	20%	0.1958 ± 0.0130
		30%	0.1931 ± 0.0109
		40%	0.1922 ± 0.0098
k-means + ML-KNN	3	103	0.1994 ± 0.0135
EM + ML-KNN	3	103	0.1953 ± 0.0120
MDI + FCM + ML-KNN	3	20%	0.1967 ± 0.0170
		30%	0.1921 ± 0.0107
		40%	**0.1899 ± 0.0118**
k-means + ML-KNN	4	103	0.2002 ± 0.0129
EM + ML-KNN	4	103	0.1984 ± 0.0120
MDI + FCM + ML-KNN	4	20%	0.1986 ± 0.0129
		30%	0.1943 ± 0.0119
		40%	0.1912 ± 0.0126
k-means + ML-KNN	5	103	0.2026 ± 0.0123
EM + ML-KNN	5	103	0.1992 ± 0.0116
MDI + FCM + ML-KNN	5	20%	0.1997 ± 0.0147
		30%	0.1941 ± 0.0113
		40%	0.1949 ± 0.0132
k-means + ML-KNN	6	103	0.2040 ± 0.0111
EM + ML-KNN	6	103	0.2017 ± 0.0124
MDI + FCM + ML-KNN	6	20%	0.1977 ± 0.0124
		30%	0.1971 ± 0.0113
		40%	0.1968 ± 0.0113
k-means + ML-KNN	7	103	0.2020 ± 0.0087
EM + ML-KNN	7	103	0.2025 ± 0.0114
MDI + FCM + ML-KNN	7	20%	0.2107 ± 0.0106
		30%	0.2030 ± 0.0106
		40%	0.2035 ± 0.0112

Table 7: Micro $F1$-measure for all models on yeast dataset. n is the number of attributes, and k is the amount of clusters.

Classification model	k	n	Yeast
k-means + ML-KNN	2	103	0.6385 ± 0.0243
EM + ML-KNN	2	103	0.6398 ± 0.0213
MDI + FCM + ML-KNN	2	20%	0.6390 ± 0.0287
		30%	0.6451 ± 0.0230
		40%	0.6481 ± 0.0240
k-means + ML-KNN	3	103	0.6332 ± 0.0246
EM + ML-KNN	3	103	0.6405 ± 0.0236
MDI + FCM + ML-KNN	3	20%	0.6372 ± 0.0335
		30%	0.6461 ± 0.0239
		40%	**0.6486 ± 0.0250**
k-means + ML-KNN	4	103	0.6314 ± 0.0251
EM + ML-KNN	4	103	0.6347 ± 0.0251
MDI + FCM + ML-KNN	4	20%	0.6378 ± 0.0277
		30%	0.6409 ± 0.0265
		40%	0.6348 ± 0.0261
k-means + ML-KNN	5	103	0.636 ± 0.0264
EM + ML-KNN	5	103	0.6309 ± 0.0225
MDI + FCM + ML-KNN	5	20%	0.6315 ± 0.0272
		30%	0.6312 ± 0.0244
		40%	0.6359 ± 0.0247
k-means + ML-KNN	6	103	0.6222 ± 0.0230
EM + ML-KNN	6	103	0.6280 ± 0.0239
MDI + FCM + ML-KNN	6	20%	0.6327 ± 0.0257
		30%	0.6384 ± 0.0248
		40%	0.6393 ± 0.0238
k-means + ML-KNN	7	103	0.6243 ± 0.0190
EM + ML-KNN	7	103	0.6258 ± 0.0227
MDI + FCM + ML-KNN	7	20%	0.6227 ± 0.0249
		30%	0.6252 ± 0.0213
		40%	0.6265 ± 0.0225

in the selected feature subset as an experimental feature subset. The Hamming loss and micro $F1$-measure results of the MLC model with the yeast dataset are shown in Tables 6 and 7. As shown in Figures 4 and 5, the Hamming loss and micro $F1$-measure demonstrate the best results when $k = 3$ with 40% of the features selected from the original data attributes. In addition, it was found that the evaluation criterion value of MLC was reduced with an increasing number of clusters.

Tables 8 and 9 show that the OCBMLC model achieved the top predictive performance (Hamming loss = 0.0879 ± 0.0048; micro $F1 = 0.7281 \pm 0.0206$) with the scene dataset. Figures 6 and 7 show that the Hamming loss and micro $F1$-measure values outperformed the "EM and ML-KNN" and "k-means and ML-KNN" models when $k = 2$ and 30% of the features of the original data attributes were selected for the scene dataset. Thus, we conclude that the proposed OCBMLC model outperforms the other classification models.

When we selected the top 30% or 40% of features using the proposed feature selection algorithm for MLC, the proposed OCBMLC model achieved the best performance because it can select features with max-dependence in consideration of the classes and interaction among features and each class. Thus, the proposed feature selection algorithm can select features with the best discrimination power. The experimental results prove that the proposed feature selection algorithm directly improves classification performance, and it is almost always better than models that use all total features from the data. The experimental results also show that the model based on overlapping clustering outperforms models based on hard clustering. In a multilabel dataset, one instance may belong to multiple labels; however, hard clustering methods attempt to assign a single instance to a single label. Therefore, such methods may not be suitable for multilabel datasets. In contrast, overlapping clustering

Table 8: Hamming loss measure for all models on scene dataset. n is the number of attributes, and k is the amount of clusters.

Classification model	k	n	Scene
k-means + ML-KNN	2	294	0.0921 ± 0.0092
EM + ML-KNN	2	294	0.0913 ± 0.0017
MDI + FCM + ML-KNN	2	20%	0.0926 ± 0.0034
		30%	**0.0879 ± 0.0048**
		40%	0.0892 ± 0.0043
k-means + ML-KNN	3	294	0.0951 ± 0.0087
EM + ML-KNN	3	294	0.0957 ± 0.0008
MDI + FCM + ML-KNN	3	20%	0.0955 ± 0.0032
		30%	0.0872 ± 0.0051
		40%	0.0908 ± 0.0043
k-means + ML-KNN	4	294	0.0948 ± 0.0098
EM + ML-KNN	4	294	0.0990 ± 0.0014
MDI + FCM + ML-KNN	4	20%	0.0973 ± 0.0040
		30%	0.0917 ± 0.0043
		40%	0.0925 ± 0.0019
k-means + ML-KNN	5	294	0.0994 ± 0.0106
EM + ML-KNN	5	294	0.1028 ± 0.0032
MDI + FCM + ML-KNN	5	20%	0.0974 ± 0.0017
		30%	0.0915 ± 0.0055
		40%	0.0949 ± 0.0108
k-means + ML-KNN	6	294	0.1001 ± 0.0073
EM + ML-KNN	6	294	0.1056 ± 0.0028
MDI + FCM + ML-KNN	6	20%	0.0970 ± 0.0015
		30%	0.0957 ± 0.0060
		40%	0.0956 ± 0.0116
k-means + ML-KNN	7	294	0.0987 ± 0.0068
EM + ML-KNN	7	294	0.1050 ± 0.0033
MDI + FCM + ML-KNN	7	20%	0.0989 ± 0.0014
		30%	0.0965 ± 0.0026
		40%	0.0975 ± 0.0026

Table 9: Micro $F1$-measure for all models on scene dataset. n is the number of attributes, and k is the amount of clusters.

Classification model	k	n	Scene
k-means + ML-KNN	2	294	0.7146 ± 0.0280
EM + ML-KNN	2	294	0.7107 ± 0.0075
MDI + FCM + ML-KNN	2	20%	0.7161 ± 0.0190
		30%	**0.7281 ± 0.0206**
		40%	0.7257 ± 0.0145
k-means + ML-KNN	3	294	0.7060 ± 0.0269
EM + ML-KNN	3	294	0.7011 ± 0.0094
MDI + FCM + ML-KNN	3	20%	0.7115 ± 0.0149
		30%	0.7235 ± 0.0227
		40%	0.7231 ± 0.0162
k-means + ML-KNN	4	294	0.7051 ± 0.0278
EM + ML-KNN	4	294	0.6864 ± 0.0023
MDI + FCM + ML-KNN	4	20%	0.7061 ± 0.0158
		30%	0.7188 ± 0.0211
		40%	0.7187 ± 0.0195
k-means + ML-KNN	5	294	0.6916 ± 0.0330
EM + ML-KNN	5	294	0.6724 ± 0.0096
MDI + FCM + ML-KNN	5	20%	0.7108 ± 0.0103
		30%	0.7010 ± 0.0305
		40%	0.7001 ± 0.0332
k-means + ML-KNN	6	294	0.6900 ± 0.0187
EM + ML-KNN	6	294	0.6606 ± 0.0080
MDI + FCM + ML-KNN	6	20%	0.6998 ± 0.0024
		30%	0.6993 ± 0.0322
		40%	0.6961 ± 0.0413
k-means + ML-KNN	7	294	0.6938 ± 0.0211
EM + ML-KNN	7	294	0.6637 ± 0.0143
MDI + FCM + ML-KNN	7	20%	0.6870 ± 0.0095
		30%	0.6958 ± 0.0147
		40%	0.6918 ± 0.0147

methods consider situations where single instances do in fact belong to multiple classes.

6. Conclusion

This paper has proposed an overlapping clustering-based MLC model that includes a feature selection phase for original datasets. We have also proposed a new multilabel feature selection algorithm that can effectively select significant features to improve classification performance. The proposed MLC framework includes an initial overlapping clustering phase. The proposed model considers the fact that multilabel data examples may not be related to a single class but may belong to multiple classes in many cases. Therefore, overlapping clustering may be more suitable for such situations. Experimental results show that the proposed model can increase predictive performance compared to a nonoverlapping clustering framework. In addition, the results demonstrate that feature selection plays an important role in classification. In future, we plan to further explore and develop a better and more robust feature selection method or overlapping clustering algorithm for MLC tasks.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 60903074 and the National High Technology Research and Development Program of China (863 Program) under Grant 2008AA01Z119.

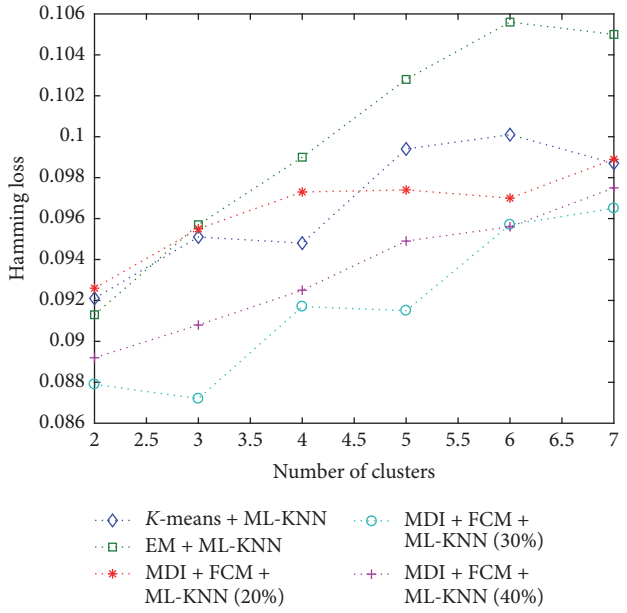

Figure 6: Hamming loss for all models and the number of clusters in scene.

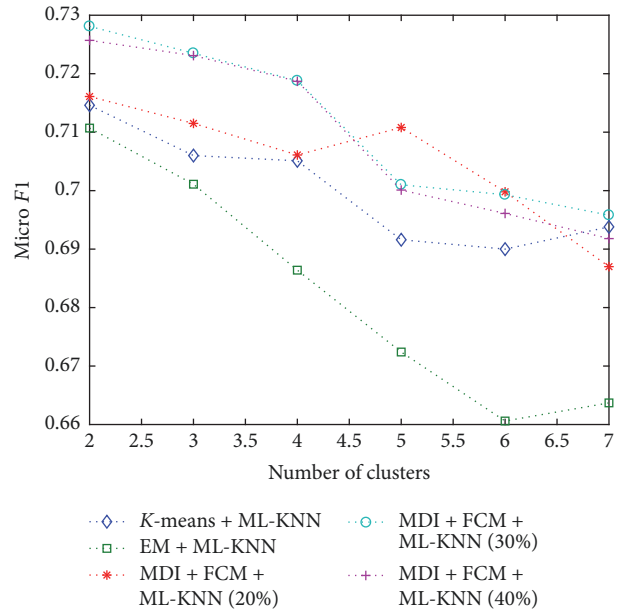

Figure 7: Micro $F1$ measure for all models and the number of clusters in scene.

References

[1] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Machine Learning*, vol. 73, no. 2, pp. 185–214, 2008.

[2] J. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognition*, vol. 48, no. 9, pp. 2761–2771, 2015.

[3] J. Lee and D. W. Kim, "Memetic feature selection algorithm for multi-label classification," *Information Sciences*, vol. 293, pp. 80–96, 2015.

[4] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 261–274, 2008.

[5] K. Feng, J. Rong, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, pp. 1719–1726, USA, June 2006.

[6] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," Ecml/pkdd Workshop on Mining Multidimensional Data, 2008.

[7] D. Mena, E. Montañés, J. R. Quevedo, and J. J. del Coz, "An overview of inference methods in probabilistic classifier chains for multilabel classification," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 6, pp. 215–230, 2016.

[8] G. Tsoumakas and I. Vlahavas, "Random k-Labelsets: An Ensemble Method for Multilabel Classification," in *Machine Learning: ECML*, 2007.

[9] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Data mining and knowledge discovery handbook," *Kybernetes*, vol. 33, no. 7, pp. 809–835, 2010.

[10] K. Brinker and E. Hullermeier, "Case-based multi-label ranking," in *Proceeding of the Conference on Artificial Intelligence (IJCAI'07)*, pp. 702–707, India, 2007.

[11] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[12] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[13] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[14] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 5, no. 4, pp. 537–550, 1994.

[15] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[16] 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, 2015.

[17] Z. C. Li, J. Liu, Y. Yang et al., "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 9, pp. 2138–2150, 2014.

[18] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5343–5355, 2015.

[19] D. Cai, X. He, and J. Han, "Spectral regression: a unified approach for Sparse Subspace Learning," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 73–82, Omaha, Neb, USA, October 2007.

[20] Z. C. Li, J. H. Tang et al., "Robust structured subspace learning for data representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 10, pp. 2085–2098, 2015.

[21] G. R. Kommu, M. Trupthi, and S. Pabboju, "A novel approach for multi-label classification using probabilistic classifiers," in *Proceedings of the 2014 International Conference on Advances in Engineering and Technology Research, ICAETR 2014*, India, August 2014.

[22] T. Guo and G. Li, "Improved Conditional Dependency Networks for Multi-label Classification," in *Proceeding of the International Conference on Measuring Technology & Mechatronics Automation*, pp. 561–565, 2015.

[23] Y. Yu, W. Pedrycz, and D. Miao, "Neighborhood rough sets based multi-label classification for automatic image annotation," *International Journal of Approximate Reasoning*, vol. 54, no. 9, pp. 1373–1387, 2013.

[24] G. Nasierding, G. Tsoumakas, and A. Z. Kouzani, "Clustering based multi-label classification for image annotation and retrieval," in *Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, SMC 2009*, pp. 4514–4519, USA, October 2009.

[25] G. Nasierding and A. Sajjanhar, "Multi-label classification with clustering for image and text categorization," in *Proceedings of the 2013 6th International Congress on Image and Signal Processing, CISP 2013*, pp. 869–874, China, December 2013.

[26] G. Nasierding, Y. Li, and A. Sajjanhar, "Robustness comparison of clustering - Based vs. non-clustering multi-label classifications for image and video annotations," in *Proceedings of the 8th International Congress on Image and Signal Processing, CISP 2015*, pp. 691–696, China, October 2015.

[27] M. A. Tahir, J. Kittler, K. Mikolajczyk et al., "Improving multilabel classification performance by using ensemble of multi-label classifiers," in *Proceedings of the International Conference on Multiple Classifier Systems, International Workshop, DBLP*, pp. 11–21, 2010.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.

[29] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.

[30] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[31] J. Nayak, B. Naik, and H. Behera, "Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014," in *Computational Intelligence in Data Mining—Volume 2*, L. C. Jain, H. S. Behera, J. K. Mandal, and D. P. Mohapatra, Eds., vol. 32 of *Smart Innovation, Systems and Technologies*, pp. 133–149, Springer, 2015.

[32] K. Zou, Z. Wang, and M. Hu, "An new initialization method for fuzzy *c*-means algorithm," *Fuzzy Optimization and Decision Making. A Journal of Modeling and Computation Under Uncertainty*, vol. 7, no. 4, pp. 409–416, 2008.

[33] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," in *Proceedings of the International Symposium on Music Information Retrieval*, pp. 325–330, September 2008.

[34] A. Elissee and J. Weston, "A kernel method for multi-labelled classification," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 681–687, 2001.

[35] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[36] G. Tsoumakas, I. Katakis, and I. Vlahavas, "A Review of Multi-Label Classification Methods," in *Proceedings of the International Conference on the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery*, pp. 99–109, 2010.