mathematics

MDPI

Article

Multi-Scale Annulus Clustering for Multi-Label Classification

Yan Liu [1], Changshun Liu [1], Jingjing Song [1,*], Xibei Yang [1,2], Taihua Xu [1,2] and Pingxin Wang [2,3]

[1] School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China
[2] Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province, Zhoushan 316022, China
[3] School of Science, Jiangsu University of Science and Technology, Zhenjiang 212100, China
* Correspondence: songjingjing@just.edu.cn; Tel.: +86-177-0517-3990

Abstract: Label-specific feature learning has become a hot topic as it induces classification models by accounting for the underlying features of each label. Compared with single-label annotations, multi-label annotations can describe samples from more comprehensive perspectives. It is generally believed that the compelling classification features of a data set often exist in the aggregation of label distribution. In this in-depth study of a multi-label data set, we find that the distance between all samples and the sample center is a Gaussian distribution, which means that the label distribution has the tendency to cluster from the center and spread to the surroundings. Accordingly, the double annulus field based on this distribution trend, named DEPT for double annulusfield and label-specific features for multi-label classification, is proposed in this paper. The double annulus field emphasizes that samples of a specific size can reflect some unique features of the data set. Through intra-annulus clustering for each layer of annuluses, the distinctive feature space of these labels is captured and formed. Then, the final classification model is obtained by training the feature space. Contrastive experiments on 10 benchmark multi-label data sets verify the effectiveness of the proposed algorithm.

Keywords: annulus model; hierarchical clustering; label-specific features; multi-label classification

MSC: 68U01

Citation: Liu, Y.; Liu, C.; Song, J.; Yang, X.; Xu, T.; Wang, P. Multi-Scale Annulus Clustering for Multi-Label Classification. *Mathematics* **2023**, *11*, 1969. https://doi.org/10.3390/math11081969

Academic Editor: Bo Wang

Received: 24 March 2023
Revised: 13 April 2023
Accepted: 17 April 2023
Published: 21 April 2023

1. Introduction

In multi-label classification tasks, multiple classification models are derived from a training set for each class label [1]. As a popular paradigm in machine learning, multi-label learning techniques have been widely employed to solve real-world problems, such as image annotation [2], where an image may convey various information; medical diagnosis [3], where the task is to identify the patient's disease from its symptoms; and sentiment analysis [4,5], where an expression can contain many emotions, etc.

To deal with the classification problem for multi-label data, one of the common approaches is to utilize dependencies between behavioral labels to induce classification models [6–8]. Although behavioral labels achieve better results in multi-label classification, this approach might be suboptimal due to ignoring the underlying features for each of its own class labels. For example, altitude-based features are preferred in distinguishing the snow and non-snow labels, while moisture-based features tend to recognize sea and non-sea labels. Therefore, label-specific learning [9–11], which extracts the underlying characteristics of each class label, has excellent research significance in multi-label classification.

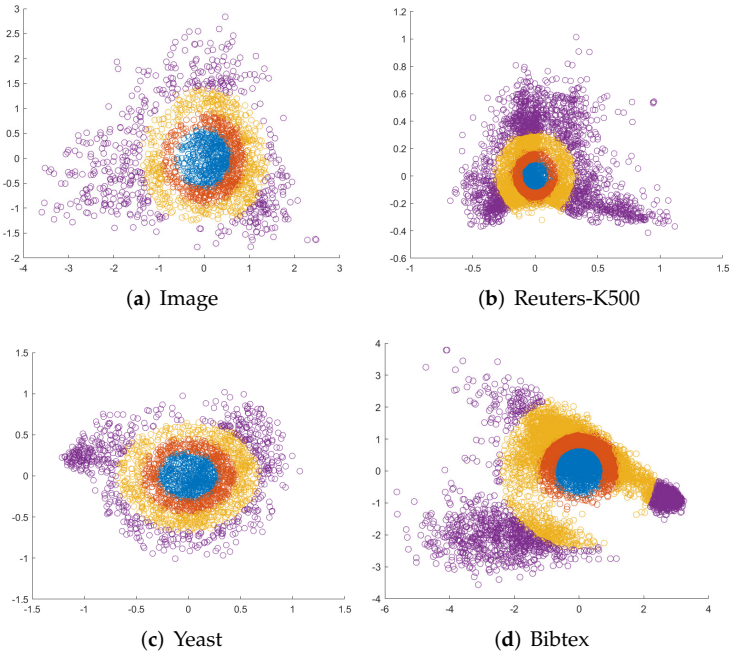
It needs to be emphasized that the degree of data aggregation intuitively reflects potentially label-specific features. Therefore, some existing methods utilize spherical random clustering, i.e., *k*-means [9,12], to extract label-specific features by evaluating the similarity between different instances for each label through Euclidean distance. The apparent advantage of the random clusters can be summarized in two points. On the one hand, the iterative process that minimizes squared error between instances and cluster

centers is easy to implement and also has low time complexity. On the other hand, favorable adaptability can be obtained, such as rapid convergence and better cluster effect. In particular, it has good scalability for extensive data conforming to the Gaussian distribution. However, there are still some limitations to this method. First, it does not view sufficient discriminative information provided by cluster centers, mainly because random clusters may be trapped in the optimal local solution [13]. Meanwhile, heterogeneous numbers of the cluster centers will further affect the results. Second, it is not suitable for discrete data set. It is also sensitive to abnormal values, i.e., it may generate cluster centers in inappropriate locations due to directly taking into account instances with significant bias.

Motivated by the above-mentioned problems, a novel label-specific feature algorithm based on the double annulus field is proposed in this paper, as shown in Figure 1, which presents the foundation of the annulus model. In other words, due to the distribution characteristics of a multi-label data set, where the majority of samples cluster around the sample center and exhibit a divergent trend towards the surrounding areas, we propose an adaptive annulus model for the distribution trend of a multi-label data set. It should be noted that the double annulus field is an improvement on the single annulus model. The single annulus model is used to construct the label-specific feature space by hierarchically extracting each layer of instances. Although significant biases in the number of positive and negative instances are mitigated through the model, classification ability for more indistinguishable cases is limited, such as the cross-distribution of instances and unbalanced label density. That is, instances within the single annulus are not enough to provide discernable information for both categories simultaneously. In response to this situation, the double annulus field is proposed. The principle of this model is to divide the heterogeneously intersecting instances with high density into different annulus fields, and annulus clustering is then carried out. One of the essential tasks of the double annulus field is to design a mapping process. The mapping process that uses principal component analysis (PCA) [14] can not only bisect each category of instances into two parts but also reduce the classification difficulty in terms of the unbalanced label density. In summary, the contributions of this paper are highlighted as follows:

1. The concept of the double annulus field for multi-label data is established, which ensures the same number of instances in each layer of annuluses and thus sufficient cluster information to be contained within each annulus.
2. A cluster strategy within the annulus model is developed, which captures the potential features for each layer of instances and thus effectively prevents information loss. To mitigate the influence of unbalanced label density, each layer of instances can be divided by mapping relationships.

The rest of this paper is organized as follows. In Section 2, several multi-label learning methods are reviewed. In Section 3, the double annulus field for exploring hierarchical label-specific features and intra-annulus clusters are both analyzed. The comparative experimental results and analyses are shown in Section 4. Finally, the conclusions are summarized in Section 5.



(**a**) Image (**b**) Reuters-K500

(**c**) Yeast (**d**) Bibtex

Figure 1. Illustrative examples of four multi-label data sets in a two-dimensional case. The annuluses are generated in the distribution trend of a multi-label data set.

2. Related Work

As a pragmatic and popular paradigm in machine learning, multi-label classification has been used intensively in recent research to study aspects such as label correlations [15,16], data streams [17,18], class imbalance [19,20], and feature selection [21,22].

Recent approaches to label correlations can be divided into three techniques by considering the order of label correlations, namely first-order [23,24], second-order [25–27], and high-order techniques [28–30]. Data streams differentiate themselves from traditional one-time scanning of all data by emphasizing the fact that the concepts contained in the data will change over time and by processing data on time using limited resources, i.e., memory and time. Class imbalance indicates an inherent attribute of multi-label data, where the number of positive training instances is generally much less than its negative counterparts. Feature selection is an effective dimension reduction technique to cope with high-dimensional multi-label data.

Differing from the above methods, label-specific features as an intuitive approach to dealing with multi-label classification problems could explore tailored features and construct feature mappings. In addition, the main idea in such an approach is that the underlying characteristics for each class label are different in their discrimination processes. Therefore, a more effective classification model can be induced through the discriminative features under each label. To achieve this goal, Zhang et al. [9] proposed the LIFT (multi-label learning with label-specific features) algorithm that obtains label-specific features by random clustering and then forms the classification models through the mapping relationships of label-specific features. Xu et al. [12] reduced redundant information on increasing dimensions of multi-label data and performed sample selection through the fuzzy rough set [31]. Zhan et al. [32] appended the ensemble clustering strategy to optimize the unstable random clustering in LIFT. From a deep learning perspective, CLIF (collaborative learning of label semantics and deep label-specific features for multi-label classification) [33] further integrated deep neural networks and label semantics [34] to guide the formation of label-specific features.

Combining label-specific features with the correlation of pairwise labels, Pei et al. [35] proposed JLSE2N (joint label-density-margin space and extreme elastic net for label-specific features), which utilizes the density of multi-label data to calculate the cosine similarity,

and then quickly forms the label-specific features based on the elastic net. Lin et al. [36] proposed MULFE (multi-label learning via label-specific feature space ensemble) to obtain maximum margin multi-label classification through ensemble learning and label correlation. Zhang et al. [37] proposed BiLabel (bilabel-specific feature generation and predictive model induction), which emphasized the generation process of label-specific features and generated the label-specific features through heuristic prototype selection. Different from the above-mentioned methods, we focus on stratification in processing multi-label data. Specifically, as shown in Figure 2, the conception of annulus division is constructed based on positive and negative instances to extract hierarchical label-specific features in the design of the double annulus field.

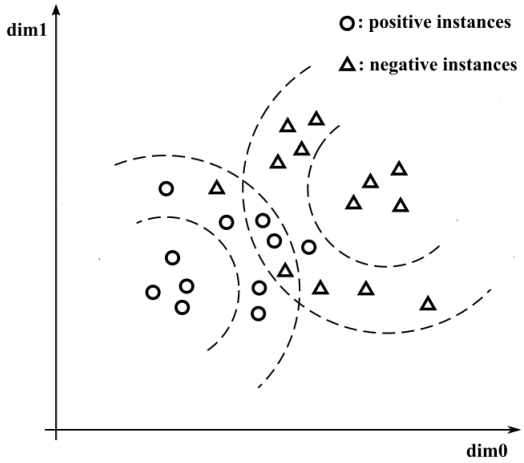


Figure 2. Illustrative example of artificial data in a two-dimensional case. Double annuluses are generated in the centers of positive instances and negative instances.

3. The Annulus Feature Space

3.1. Notations

First of all, the following notations are utilized in this paper. Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid 1 \leq i \leq m\}$ represent the multi-label training set with m instances, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector, $\boldsymbol{y}_i \in \{0,1\}^{1\times q}$ is the possible set of q labels associated with $\boldsymbol{x}_i$, and $l = \{l_1, l_2, ..., l_q\}$ denotes the label space with q labels. Formally, $y_{ik} = 1$ represents $l_k \in y_{ik}$, and $y_{ik} = 0$ represents $l_k \notin y_{ik}$.

3.2. Single Annulus Approach

The single annulus approach aims to construct a label-specific feature space by hierarchically capturing the underlying features of each label. According to each label, the positive and negative samples are separated into two different parts, i.e., POS_k and NEG_k, respectively, based on whether it contains the kth label.

To mitigate the class imbalance problem, i.e., $|POS_k| \ll |NEG_k|$, the single annulus centers are selected for a smaller number of samples PON_k:

$$PON_k = \begin{cases} POS_k, |POS_k| \leq |NEG_k|; \\ NEG_k, |POS_k| > |NEG_k|, \end{cases} \tag{1}$$

where $|\cdot|$ denotes the cardinality of the set.

Based on the selected training set PON_k, the center of the annuluses can be determined as $C_k = \text{mean}(PON_k)$. To ensure that each annulus contains the same amount of information, the single annulus approach supposes the instances are divided into r annuluses under each class label, and the instances $\mathcal{Q}_k^j$ contained in the jth annulus are also determined accordingly.

For constructing a single annulus clustering feature space, the instances with the shortest sum of distances to others are taken as the cluster center of each layer. Therefore, the positive and negative cluster centers for the jth layer of the annuluses can be respectively expressed as follows:

$$\begin{aligned} pc_j^k &= \left\{ \boldsymbol{x}_i \mid \forall \boldsymbol{x} \in \mathcal{Q}_{k+}^j, \sum_{q=1}^{\alpha} d(\boldsymbol{x}_i, \boldsymbol{x}_q) \leq \sum_{q=1}^{\alpha} d(\boldsymbol{x}, \boldsymbol{x}_q) \right\}, \\ nc_j^k &= \left\{ \boldsymbol{x}_i \mid \forall \boldsymbol{x} \in \mathcal{Q}_{k-}^j, \sum_{q=1}^{\alpha} d(\boldsymbol{x}_i, \boldsymbol{x}_q) \leq \sum_{q=1}^{\alpha} d(\boldsymbol{x}, \boldsymbol{x}_q) \right\}. \end{aligned} \tag{2}$$

Conceptually, we can already obtain the potential features in each layer of annuluses through single annulus clustering. Those features provide prototypes for constructing a single annulus clustering feature space. Here, mapping $\varphi_k : \mathcal{D} \to SCFS_k$ from the original training set to a single annulus clustering feature space can be defined as:

$$\varphi_k(\boldsymbol{x}_i) = [d\left(\boldsymbol{x}_i, pc_1^k\right), \ldots, d\left(\boldsymbol{x}_i, pc_r^k\right), d\left(\boldsymbol{x}_i, nc_1^k\right), \ldots, d\left(\boldsymbol{x}_i, nc_r^k\right)], \tag{3}$$

in which $d(\cdot, \cdot)$ denotes the Euclidean distance between two instances.

Finally, a family of l classification models can be trained by the single annulus feature mapping φ_k for the kth class label. Here, for each class label $l_k \in \mathcal{Y}$, a new binary training set $\mathcal{G}_k$ is created from the original training set $\mathcal{T}$, and the mapping of φ_k can be set as follows:

$$\mathcal{G}_k = \{(\varphi_k(\boldsymbol{x}_i), y_{ik}) \mid \boldsymbol{x}_i \in \mathcal{D}, y_{ik} \in \mathbf{Y}_i\}. \tag{4}$$

Correspondingly, a classification model $\mathcal{V}_k : SCFS_k \to \mathbb{R}$ for the kth class label can be induced by utilizing any binary learner [9]. Therefore, an unseen example $\boldsymbol{x}'$, which is associated with label set L', can be predicted as: $Y' = \left\{ L_k' \mid \mathcal{V}_k(\varphi_k(\boldsymbol{x}')) > 0, 1 \leq k \leq l \right\}$.

3.3. Double Annulus Field

The sample distribution tendency for each label represents an essential and intuitive connection with feature extraction. In this section, we attempt to construct the double annulus field based on the distribution tendency, where the classification difficulty of heterogeneous samples can be mitigated through the hierarchical division of double annuluses.

Before stratifying positive and negative samples, it is necessary to determine the two centers of the double annulus model. Through a large number of observations on a multi-label data set, it is obvious that most of the samples are clustered in the sample center of positive and negative samples. To preferentially partition the densely-distributed samples, the mean of POS_k is used to indicate the center of the positive samples, i.e., C_p^k uses the mean of POS_k to indicate the center of the positive samples, while C_n^k uses the mean of NEG_k to indicate the center of the negative samples. In addition, POS_k contains the instances $\boldsymbol{x}_i^+$ with $y_{ik} = 1$, and NEG_k contains the instances $\boldsymbol{x}_i^-$ with $y_{ik} = 0$. The numbers of POS_k and NEG_k for the kth label are defined by m_k^+ and m_k^-, respectively. In further measuring the discrete degree of the samples, the distance of the heterogeneous samples from the center point is calculated for the kth label:

$$\begin{aligned} G_k^+ &= [d(\boldsymbol{x}_1^+, C_p^k), d(\boldsymbol{x}_2^+, C_p^k), \ldots, d(\boldsymbol{x}_{m_k^+}^+, C_p^k)], \\ G_k^- &= [d(\boldsymbol{x}_1^-, C_p^k), d(\boldsymbol{x}_2^-, C_p^k), \ldots, d(\boldsymbol{x}_{m_k^-}^-, C_p^k)], \end{aligned} \tag{5}$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance between two instances and G_k^+ represents the set of Euclidean distances between positive samples and C_p^k. In addition, SG_k^+ is used to denote an ascending sort of G_k^+, while SG_k^- is used to denote an ascending sort of G_k^-.

It is well known that the cluster center points generated by random clustering will be affected by densely-distributed samples without obvious discrimination. Therefore, it is essential to ensure that the same amount of information is contained in each layer of the annulus. The extraction of latent features will not be ignored due to the extreme aggregation of samples. α represents the number of instances in each annulus, as long as the number of annuluses r is specified. Then, the maximum amount of information contained in each layer of positive samples can be defined as:

$$\alpha_k^+ = \left\lceil \frac{1}{r} \cdot |POS_k| \right\rceil. \tag{6}$$

Similarly, the maximum amount of information contained in each layer of negative samples can be defined as $\alpha_k^- = \left\lceil \frac{1}{r} \cdot |NEG_k| \right\rceil$.

In the next step, the positive samples are stratified according to the ascending distance from the center point SG_k^+, which allows us to obtain instances with α_k^+ stride separately according to the sorted order and assemble them into each layer of the annulus. More specifically, the maximum positive samples contained in the jth annulus for the kth label can be defined as:

$$\mathcal{R}_{k+}^j = \underbrace{[\, \boldsymbol{x}_u, \boldsymbol{x}_{u+1}, \cdots, \boldsymbol{x}_{v-1}, \boldsymbol{x}_v \,]}_{v-u+1=\alpha_k^+ \times j} \in SG_k^+. \tag{7}$$

Similarly, we can also obtain the maximum negative samples contained in the jth annulus for the kth label $\mathcal{R}_{k-}^j$. In other words, $\mathcal{R}_{k+}^j$ and $\mathcal{R}_{k-}^j$ together form the double annulus field.

3.4. Intra-Annulus Clustering

Observing the distribution of the multi-label data set, we can see that, in addition to the positive and negative samples concentrated in C_p^k and C_n^k, there are also more samples at the junction of the positive and negative samples. To further reduce the difficulty of classification, the boundary samples between positive and negative samples are divided into left and right parts relative to the sample centers by PCA, i.e., P_k^+ (relative left division of instances in POS_k), P_k^- (relative right division of instances in POS_k), N_k^+ (relative left division of instances in NEG_k), and N_k^- (relative right division of instances in NEG_k). Correspondingly, P_k^+, P_k^-, N_k^+, and N_k^- are defined as follows:

$$\begin{aligned} P_k^+ &= \left\{ \boldsymbol{x}_i \mid \pi_1(\boldsymbol{x}_i) < \pi_1\left(C_p^k\right) \right\}, \\ P_k^- &= \left\{ \boldsymbol{x}_i \mid \pi_1(\boldsymbol{x}_i) \geq \pi_1\left(C_p^k\right) \right\}, \\ N_k^+ &= \left\{ \boldsymbol{x}_i \mid \pi_1(\boldsymbol{x}_i) < \pi_1\left(C_n^k\right) \right\}, \\ N_k^- &= \left\{ \boldsymbol{x}_i \mid \pi_1(\boldsymbol{x}_i) \geq \pi_1\left(C_n^k\right) \right\}, \end{aligned} \tag{8}$$

where $\pi_n(\cdot)$ uses PCA to reduce the features to n-dimensions.

Since the same amount of information is guaranteed to be contained in each annulus, there are potential features that cannot be ignored. Therefore, $2r$ cluster centers are generated according to the left and right division of relative positions in each layer of the annulus, and the Euclidean distance between each cluster center and other samples is guaranteed to be the smallest. For the convenience of the following expressions, we take "$\cdot$" as the state

parameter representing the two symbols of "±". The intra-annulus cluster centers within the jth layer of the double annulus field for each label can be defined as:

$$\begin{aligned} pc_j^{k\cdot} &= \left\{ \boldsymbol{x}_i \mid \forall \boldsymbol{x} \in P_k^{\cdot} \bigcap \mathcal{R}_{k\cdot}^j, \sum_{q=1}^{p\cdot} d(\boldsymbol{x}_i, \boldsymbol{x}_q) \leq \sum_{q=1}^{p\cdot} d(\boldsymbol{x}, \boldsymbol{x}_q) \right\}, \\ nc_j^{k\cdot} &= \left\{ \boldsymbol{x}_i \mid \forall \boldsymbol{x} \in N_k^{\cdot} \bigcap \mathcal{R}_{k\cdot}^j, \sum_{q=1}^{n\cdot} d(\boldsymbol{x}_i, \boldsymbol{x}_q) \leq \sum_{q=1}^{n\cdot} d(\boldsymbol{x}, \boldsymbol{x}_q) \right\}, \end{aligned} \tag{9}$$

where $p\cdot = |P_k^{\cdot}| - |\mathcal{R}_k^{\cdot}|$, $n\cdot = |N_k^{\cdot}| - |\mathcal{R}_k^{\cdot}|$.

In this way, sufficient distinguishable information is provided for forming the label-specific features through the mapping relationship between the instances and the annulus cluster centers. Those features provide prototypes for constructing label-specific feature spaces concerning the double annulus field. Meanwhile, a mapping $\varphi_k^{'} : \mathcal{D} \to DEPT_k$ from the original training set to label-specific feature space with respect to the double annulus field can be defined as:

$$\varphi_k(\boldsymbol{x}_i) = \left[d\left(\boldsymbol{x}_i, pc_1^{k+}\right), \ldots, d\left(\boldsymbol{x}_i, pc_{2r}^{k-}\right), d\left(\boldsymbol{x}_i, nc_1^{k+}\right), \ldots, d\left(\boldsymbol{x}_i, nc_{2r}^{k-}\right) \right]. \tag{10}$$

3.5. Classification

Therefore, a family of l classification models can be trained by the double annulus field's label-specific feature space φ_k for the kth class label. Here, for each class label $l_k \in \boldsymbol{y}$, a new binary training set h_k is created from the original training set, and the mapping of φ_k can be set as follows [9]:

$$h_k = \{ (\varphi_k(\boldsymbol{x}_i), y_{ik}) \mid \boldsymbol{x}_i \in \mathcal{D}, y_{ik} \in \boldsymbol{y}_i \}. \tag{11}$$

Correspondingly, a classification model $\mathcal{V}_k : DEPT_k \to \mathbb{R}$ for the kth class label can be induced by utilizing any binary learner [9]. Therefore, an unseen example $\boldsymbol{x}'$, which is associated with label set L', can be predicted as [9]:

$$Y' = \left\{ L_k' \mid \mathcal{V}_k(\varphi_k(\boldsymbol{x}')) > 0, 1 \leq k \leq l \right\}; \tag{12}$$

in other words, from Equations (11) and (12), each classifier h_k with reference to the kth class label can be regarded as the composition of $\mathcal{V}_k$ and φ_k. Namely, $h_k(\boldsymbol{x}') = [\mathcal{V}_k \circ \varphi_k](\boldsymbol{x}') = \mathcal{V}_k(\varphi_k(\boldsymbol{x}'))$.

4. Experiments

4.1. Data Set

The details of the data set used in the experiments are shown in Table 1. In addition, we use Card, DL, and Den [1,29] to represent the average of label cardinality, the number of different label sets, and label density, respectively.

4.2. Evaluation Metrics

Compared with traditional single-label criteria, the performance of each evaluation metric is somewhat more complicated, as each instance is related to different labels concurrently. Therefore, Hamming loss (HL), average precision (AP), macro-averaging AUC (AUC), one-error (OE), coverage (CV), and ranking loss (RL) are selected in this paper. Given a test set $\mathcal{T} = \{ (\boldsymbol{x}_i, \mathbf{Y}_i) \mid 1 \leq i \leq m_t \}$, where $\mathbf{Y}_i \in \mathcal{Y}$ is the objectively true label subset, $P_x^{(i)} \in \mathcal{Y}$ is a predicted label vector for the ith instance, and $\mu_j^{(i)}$ is a confidence score that denotes the degree of $\boldsymbol{x}_i$ that belongs to the label; the details of six evaluation metrics are given as follows.

Table 1. Characteristics of the 10 multi-label data sets.

Data Set	m	d	q	Type	Domain	Label Features		
						Card	DL	Den
Birds	645	260	19	both	audio	1.014	133	0.053
CHD-49	555	49	6	numeric	medicine	2.580	34	0.430
Emotions	593	72	6	numeric	music	1.869	27	0.311
Flags	194	19	7	both	images	3.392	54	0.485
Images	2000	294	5	numeric	images	1.236	20	0.247
Medical	978	1449	45	nominal	text	1.245	84	0.028
Reuters-K500	6000	500	103	numeric	text	1.462	811	0.014
Scene	2407	294	6	numeric	images	1.074	15	0.179
WaterQuality	1060	16	14	numeric	chemistry	5.073	825	0.362
Yeast	2417	103	14	numeric	biology	4.237	198	0.303

Hamming loss [38] evaluates the number of misclassified instance-label pairs:

$$HL = \frac{1}{m_t}\sum_{i=1}^{m_t}|P_x^{(i)} \oplus Y_i|, \tag{13}$$

where $\oplus$ is the XOR operator.

Macro-averaging AUC [1] evaluates the average AUC value of each label:

$$AUC = \frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|}\frac{|\left\{(\mu_1,\mu_2) \mid \mu_1^{(i)} \geq \mu_2^{(i)}, (\mu_1,\mu_2) \in \mathcal{Z}_j \times \bar{\mathcal{Z}}_j\right\}|}{|\mathcal{Z}_j||\bar{\mathcal{Z}}_j|}. \tag{14}$$

One-error [39] evaluates whether the top-ranked predicted value of the samples is in the objectively true label subset:

$$OE = \frac{1}{m_t}\sum_{i=1}^{m_t}[\![[\arg\max_{y\in\mathcal{Y}} \mu^{(i)}] \in Y_i]\!], \tag{15}$$

where $[\![\cdot]\!]$ is a function for logical judgment.

Ranking loss [40] evaluates the fraction of reversely-ordered label pairs. Namely, ranking and label relevance are negatively correlated:

$$RL = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{1}{|Y_i||\overline{Y}_i|}|\left\{(\mu_1,\mu_2) \mid \mu_1^{(i)} \leq \mu_2^{(i)}, (\mu_1,\mu_2) \in Y_i \times \overline{Y}_i\right\}|, \tag{16}$$

where $\overline{Y}_i$ is the complementary set of Y_i.

Coverage [41] evaluates the average of steps needed to move down the ranked label list that covers all the related labels of the samples:

$$CV = \frac{1}{m_t}\sum_{i=1}^{m_t} rank_{\mu^{(i)}} - 1. \tag{17}$$

Average precision [42] evaluates the average fraction of relevant labels ranked higher than a particular label $y \in Y_i$:

$$AP = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{1}{|Y_i|}\sum_{y_i\in Y_i}\frac{|\left\{y' \mid rank_{\mu_1^{(i)}} \leq rank_{\mu_2^{(i)}}, y' \in Y_i\right\}|}{rank_{\mu^{(i)}}}. \tag{18}$$

4.3. Comparative Studies

In this section, the proposed DEPT algorithm is compared with the following six algorithms to verify their predictive performances. The only parameter of r is set to 12, according to the subsequent discussion.

- MLNB [43]: This algorithm adopts the traditional naive Bayes classifiers to deal with multi-label instances, and the parameter is the default value of 0.3.
- MULFE [36]: It combines ensemble learning with label correlation to construct a multi-label-specific feature space. The parameter of λ is set to 0.1 in this paper.
- LIFT [9]: It transforms a multi-label into a single-label problem, and then the particular features of each label are used to form a label-specific feature space. In this paper, the ratio parameter ϵ is set to 0.1.
- FRS-LIFT [12]: Based on label-specific features, this algorithm selects samples and reduces the dimension of the feature space by utilizing the idea of rough sets. The parameter setting is the same as LIFT.
- ML-KNN [40]: This algorithm is derived from the traditional k-nearest neighbor (k-NN) algorithm, in which the number of nearest neighbors is set to 10, and the smoothing parameter is set to 1.
- SCFS: Hierarchical label-specific features are extracted from the perspective of a single annulus. The parameter setting is the same as DEPT.

  For a fair comparison, LIBSVM [44] has utilized the linear kernel for training and predicting the induced binary classification model. Moreover, 10-fold cross-validation is used for each compared algorithm. Namely, the data set is divided into 10 groups of the same size; each algorithm is trained repeatedly 10 times on 9 random groups, and one group is used for testing.

Tables 2–4 report the detailed experimental results in terms of Hamming loss, average precision, macro-averaging AUC, one-error, ranking loss, and coverage, respectively. Results are compared by mean ± std, and the best performance for each data set is shown in bold. Additionally, "↑" and "↓" represent "the larger, the better" and "the smaller, the better", respectively.

Table 2. Experimental results of each compared algorithm (mean ± std).

Comparison Algorithm	Average Precision ↑						
	DEPT	SCFS	LIFT	MLKNN	MLNB	FRSLIFT	MULFE
Birds	**0.750 ± 0.044**	0.736 ± 0.041	0.608 ± 0.030	0.587 ± 0.022	0.579 ± 0.059	0.674 ± 0.023	0.733 ± 0.076
CHD-49	**0.814 ± 0.024**	0.806 ± 0.026	0.805 ± 0.034	0.773 ± 0.022	0.789 ± 0.035	0.813 ± 0.025	0.813 ± 0.033
Emotions	0.819 ± 0.029	0.815 ± 0.021	**0.829 ± 0.007**	0.814 ± 0.017	0.768 ± 0.026	0.803 ± 0.032	0.812 ± 0.030
Flags	**0.828 ± 0.033**	0.822 ± 0.026	0.820 ± 0.021	0.796 ± 0.070	0.802 ± 0.043	0.812 ± 0.030	0.819 ± 0.039
Images	0.805 ± 0.028	0.791 ± 0.018	0.810 ± 0.030	0.779 ± 0.029	0.756 ± 0.030	**0.831 ± 0.023**	0.823 ± 0.020
Medical	**0.868 ± 0.035**	0.825 ± 0.034	0.862 ± 0.005	0.812 ± 0.029	0.777 ± 0.007	0.867 ± 0.018	0.859 ± 0.044
Reuters-K500	**0.647 ± 0.014**	0.617 ± 0.013	0.645 ± 0.002	0.627 ± 0.019	0.592 ± 0.018	0.634 ± 0.019	0.639 ± 0.016
Scenes	0.878 ± 0.015	0.865 ± 0.014	0.886 ± 0.005	0.860 ± 0.011	0.839 ± 0.014	**0.896 ± 0.016**	0.888 ± 0.017
WaterQuality	**0.680 ± 0.016**	0.668 ± 0.036	0.673 ± 0.014	0.652 ± 0.009	0.641 ± 0.006	0.679 ± 0.026	0.663 ± 0.023
Yeast	**0.782 ± 0.019**	0.760 ± 0.016	0.766 ± 0.006	0.763 ± 0.021	0.741 ± 0.006	0.778 ± 0.018	0.768 ± 0.018
Average	**0.787**	0.770	0.770	0.746	0.729	0.779	0.782

Table 2. *Cont.*

Comparison algorithm	Hamming loss ↓						
	DEPT	SCFS	LIFT	MLKNN	MLNB	FRSLIFT	MULFE
Birds	0.045 ± 0.008	0.049 ± 0.008	**0.044 ± 0.007**	0.990 ± 0.001	0.971 ± 0.004	0.048 ± 0.009	0.050 ± 0.005
CHD-49	**0.262 ± 0.032**	0.300 ± 0.016	0.278 ± 0.023	0.322 ± 0.035	0.289 ± 0.033	0.280 ± 0.033	0.283 ± 0.020
Emotions	0.186 ± 0.020	0.194 ± 0.014	**0.177 ± 0.013**	0.785 ± 0.010	0.834 ± 0.026	0.187 ± 0.020	0.180 ± 0.020
Flags	**0.250 ± 0.049**	0.270 ± 0.029	0.280 ± 0.027	0.674 ± 0.041	0.664 ± 0.037	0.261 ± 0.018	0.274 ± 0.012
Images	0.176 ± 0.014	0.197 ± 0.012	**0.151 ± 0.013**	0.169 ± 0.008	0.199 ± 0.007	0.154 ± 0.016	0.157 ± 0.010
Medical	**0.010 ± 0.001**	0.015 ± 0.002	0.012 ± 0.002	0.984 ± 0.001	0.984 ± 0.001	0.013 ± 0.001	0.014 ± 0.002
Reuters-K500	**0.011 ± 0.001**	0.011 ± 0.000	**0.011 ± 0.000**	0.996 ± 0.000	0.963 ± 0.000	0.022 ± 0.000	0.013 ± 0.001
Scenes	0.089 ± 0.008	0.100 ± 0.005	0.076 ± 0.010	0.885 ± 0.007	0.881 ± 0.003	0.076 ± 0.015	**0.072 ± 0.006**
WaterQuality	**0.296 ± 0.009**	0.307 ± 0.011	0.297 ± 0.025	0.839 ± 0.002	0.839 ± 0.005	0.300 ± 0.027	0.307 ± 0.015
Yeast	0.191 ± 0.011	0.204 ± 0.009	0.190 ± 0.009	0.190 ± 0.008	0.205 ± 0.008	**0.189 ± 0.007**	0.196 ± 0.010
Average	**0.151**	0.165	0.152	0.683	0.683	0.153	0.155

Table 3. Experimental results of each comparison algorithm (mean ± std).

Comparison Algorithm	Macro-Averaging AUC ↑						
	DEPT	SCFS	LIFT	MLKNN	MLNB	FRSLIFT	MULFE
Birds	0.789 ± 0.045	0.704 ± 0.069	0.807 ± 0.023	**0.851 ± 0.048**	0.780 ± 0.032	0.708 ± 0.065	0.742 ± 0.045
CHD-49	**0.665 ± 0.069**	0.597 ± 0.050	0.640 ± 0.050	0.501 ± 0.061	0.635 ± 0.032	0.591 ± 0.062	0.652 ± 0.043
Emotions	0.840 ± 0.029	0.836 ± 0.022	**0.851 ± 0.014**	0.847 ± 0.019	0.814 ± 0.013	0.849 ± 0.019	0.843 ± 0.012
Flags	**0.722 ± 0.069**	0.691 ± 0.066	0.700 ± 0.042	0.686 ± 0.077	0.661 ± 0.077	0.707 ± 0.017	0.721 ± 0.022
Images	**0.832 ± 0.019**	0.795 ± 0.015	0.831 ± 0.013	0.831 ± 0.017	0.802 ± 0.018	0.829 ± 0.030	0.829 ± 0.018
Medical	**0.905 ± 0.048**	0.872 ± 0.038	0.892 ± 0.027	0.850 ± 0.033	0.724 ± 0.047	0.903 ± 0.023	0.894 ± 0.035
Reuters-K500	**0.809 ± 0.013**	0.790 ± 0.016	0.804 ± 0.030	0.777 ± 0.049	0.692 ± 0.027	0.796 ± 0.035	0.792 ± 0.022
Scenes	**0.935 ± 0.010**	0.920 ± 0.005	0.928 ± 0.008	0.924 ± 0.016	0.909 ± 0.013	0.929 ± 0.014	0.921 ± 0.011
WaterQuality	0.678 ± 0.010	0.662 ± 0.025	0.701 ± 0.023	**0.702 ± 0.010**	0.664 ± 0.012	0.692 ± 0.031	0.656 ± 0.014
Yeast	**0.729 ± 0.023**	0.638 ± 0.023	0.681 ± 0.032	0.702 ± 0.003	0.675 ± 0.030	0.623 ± 0.025	0.683 ± 0.015
Average	**0.790**	0.750	0.784	0.767	0.736	0.763	0.773
Comparison algorithm	One-error ↓						
	DEPT	SCFS	LIFT	MLKNN	MLNB	FRSLIFT	MULFE
Birds	0.668 ± 0.060	**0.657 ± 0.048**	0.689 ± 0.053	0.675 ± 0.049	0.710 ± 0.038	0.705 ± 0.038	0.694 ± 0.038
CHD-49	**0.207 ± 0.056**	0.236 ± 0.053	0.210 ± 0.071	0.219 ± 0.035	0.235 ± 0.054	0.234 ± 0.085	0.232 ± 0.061
Emotions	**0.204 ± 0.075**	0.250 ± 0.046	0.212 ± 0.026	0.254 ± 0.017	0.333 ± 0.060	0.275 ± 0.060	0.232 ± 0.070
Flags	**0.203 ± 0.066**	0.231 ± 0.084	0.248 ± 0.074	0.263 ± 0.105	0.234 ± 0.129	0.232 ± 0.120	0.232 ± 0.139
Images	0.316 ± 0.054	0.401 ± 0.029	0.285 ± 0.056	0.343 ± 0.049	0.372 ± 0.045	**0.253 ± 0.035**	0.264 ± 0.040
Medical	0.194 ± 0.048	0.212 ± 0.045	0.174 ± 0.011	0.242 ± 0.033	0.435 ± 0.056	**0.163 ± 0.027**	0.196 ± 0.042
Reuters-K500	**0.420 ± 0.018**	0.454 ± 0.018	0.421 ± 0.010	0.442 ± 0.024	0.474 ± 0.036	0.465 ± 0.027	0.453 ± 0.045
Scenes	0.222 ± 0.023	0.255 ± 0.025	0.195 ± 0.025	0.240 ± 0.014	0.261 ± 0.018	**0.185 ± 0.029**	0.195 ± 0.027
WaterQuality	0.286 ± 0.044	0.314 ± 0.044	0.327 ± 0.029	**0.277 ± 0.024**	0.340 ± 0.028	0.318 ± 0.079	0.290 ± 0.032
Yeast	**0.221 ± 0.030**	0.244 ± 0.026	0.228 ± 0.010	0.222 ± 0.038	0.240 ± 0.007	0.238 ± 0.023	0.229 ± 0.023
Average	**0.294**	0.325	0.299	0.318	0.363	0.307	0.302

Table 4. Experimental results of each comparison algorithm (mean ± std).

Comparison Algorithm	Ranking Loss ↓						
	DEPT	SCFS	LIFT	MLKNN	MLNB	FRSLIFT	MULFE
Birds	**0.170 ± 0.050**	0.222 ± 0.049	0.173 ± 0.033	0.153 ± 0.039	0.172 ± 0.049	0.206 ± 0.028	0.193 ± 0.065
CHD-49	0.234 ± 0.052	0.238 ± 0.028	**0.211 ± 0.028**	0.260 ± 0.066	0.254 ± 0.023	0.223 ± 0.036	0.238 ± 0.038
Emotions	0.148 ± 0.031	0.161 ± 0.023	0.149 ± 0.014	0.151 ± 0.021	0.184 ± 0.013	**0.140 ± 0.021**	0.148 ± 0.009
Flags	**0.206 ± 0.041**	0.209 ± 0.031	0.208 ± 0.014	0.222 ± 0.071	0.212 ± 0.022	0.217 ± 0.029	0.220 ± 0.019
Images	0.167 ± 0.022	0.207 ± 0.016	**0.140 ± 0.016**	0.182 ± 0.024	0.206 ± 0.025	0.142 ± 0.023	0.143 ± 0.011
Medical	**0.021 ± 0.020**	0.031 ± 0.013	0.024 ± 0.012	0.037 ± 0.010	0.055 ± 0.006	0.029 ± 0.004	0.032 ± 0.009
Reuters-K500	**0.040 ± 0.004**	0.067 ± 0.003	0.050 ± 0.003	0.061 ± 0.002	0.085 ± 0.004	0.483 ± 0.003	0.103 ± 0.005
Scenes	0.074 ± 0.009	0.086 ± 0.009	0.064 ± 0.003	0.079 ± 0.009	0.097 ± 0.013	**0.059 ± 0.010**	0.063 ± 0.011
WaterQuality	0.279 ± 0.013	0.295 ± 0.022	0.266 ± 0.017	**0.256 ± 0.004**	0.294 ± 0.011	0.271 ± 0.025	0.296 ± 0.025
Yeast	**0.147 ± 0.016**	0.175 ± 0.014	0.164 ± 0.012	0.151 ± 0.015	0.178 ± 0.012	0.153 ± 0.014	0.163 ± 0.007
Average	0.149	0.169	**0.145**	0.155	0.174	0.192	0.160
Comparison algorithm	Coverage ↓						
	DEPT	SCFS	LIFT	MLKNN	MLNB	FRSLIFT	MULFE
Birds	**0.102 ± 0.045**	0.114 ± 0.043	0.123 ± 0.045	0.113 ± 0.018	0.119 ± 0.059	0.139 ± 0.012	0.133 ± 0.052
CHD-49	**0.433 ± 0.026**	0.451 ± 0.021	0.450 ± 0.017	0.497 ± 0.030	0.454 ± 0.003	0.444 ± 0.010	0.453 ± 0.020
Emotions	**0.284 ± 0.023**	0.301 ± 0.040	0.290 ± 0.023	0.294 ± 0.029	0.315 ± 0.004	0.289 ± 0.024	0.301 ± 0.020
Flags	**0.513 ± 0.033**	0.526 ± 0.045	0.514 ± 0.028	0.536 ± 0.045	0.539 ± 0.037	0.576 ± 0.022	0.552 ± 0.024
Images	0.189 ± 0.014	0.220 ± 0.020	**0.164 ± 0.008**	0.192 ± 0.018	0.224 ± 0.018	0.172 ± 0.023	0.166 ± 0.017
Medical	**0.040 ± 0.023**	0.059 ± 0.016	0.042 ± 0.018	0.060 ± 0.013	0.313 ± 0.034	0.048 ± 0.009	0.055 ± 0.021
Reuters-K500	0.076 ± 0.008	0.084 ± 0.005	0.073 ± 0.006	0.090 ± 0.004	0.143 ± 0.005	**0.071 ± 0.006**	0.095 ± 0.003
Scenes	0.075 ± 0.008	0.086 ± 0.008	**0.065 ± 0.007**	0.080 ± 0.010	0.299 ± 0.016	0.067 ± 0.010	0.084 ± 0.007
WaterQuality	0.657 ± 0.018	0.675 ± 0.016	0.638 ± 0.034	**0.620 ± 0.013**	0.666 ± 0.007	0.639 ± 0.039	0.669 ± 0.023
Yeast	0.434 ± 0.016	0.443 ± 0.015	0.455 ± 0.023	**0.428 ± 0.010**	0.460 ± 0.018	0.475 ± 0.010	0.474 ± 0.016
Average	**0.280**	0.296	0.281	0.291	0.353	0.292	0.298

To analyze the statistical significance of each compared result, the Friedman test [45] with a significance level of 0.05 is employed; this is a well-known statistical test and has been widely utilized for statistically comparative studies over several data sets. Table 5 summarizes the Friedman statistics (F_F) and the corresponding critical value of each evaluation metric. As shown in Table 5, the null hypothesis that all the compared algorithms have equal performance is rejected for each evaluation metric. In addition, the Bonferroni–Dunn test [46] is used as a post-hoc test to further explore the relative performance of each compared algorithm by treating DEPT as the control algorithm. Accordingly, the significant difference between the average ranks of compared algorithms can be distinguished from others by at least one critical difference (CD):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{19}$$

where k denotes the number of compared algorithms, and N denotes the number of data sets. For the Bonferroni–Dunn test, $q_\alpha = 2.638$ at significance level $\alpha = 0.05$, and thus CD = 2.5486 ($k = 7$, $N = 10$).

To intuitively observe the relative performance of DEPT and other comparison algorithms, Figure 3 shows the CD diagrams in terms of each evaluation metric, where the compared algorithm marked on the left axis has lower performance than the right one. Otherwise, any compared algorithm that is not significantly different from DEPT (within one CD) is connected with a black line.

Table 5. Summary of the Friedman statistics F_F (k = 7, N= 10) and the critical value in terms of each evaluation metric (k: number of comparison algorithms; N: number of data sets).

Evaluation Metric	F_F	Critical Value (α = 0.05)
Average precision	17.8657	2.2720
Macro-averaging AUC	7.0714	
Coverage	6.4649	
Ranking loss	5.2534	
Hamming loss	13.8468	
One-error	6.0898	

Based on the reported experimental results, the following conclusions can be drawn:

1. As shown in Figure 3, DEPT significantly outperforms MLNB in terms of each evaluation metric. Furthermore, DEPT differs significantly in performance from other algorithms in more than 38% of cases. That is, in terms of the average precision metric, DEPT is significantly different from MLNB and MLKNN. In terms of the macro-averaging AUC metric, DEPT is significantly different from MLNB and SCFS. In terms of the coverage metric, DEPT is significantly different from MLNB, MULFE, and SCFS. In terms of the Hamming loss metric, DEPT has is statistically significantly different from MLNB, MLKNN, and SCFS. In terms of the one-error metric, DEPT has a statistically significant difference between MLNB and SCFS. Therefore, the DEPT algorithm is statistically superior to others in more than 38% of cases.
2. As shown in Figure 3, DEPT shows statistically superior or at least comparable performance against SCFS in each evaluation metric. Although the performance of SCFS is similar to MLNB due to the stratified feature space, SCFS uses a single annulus model that can not appropriately solve the cross-distribution of instances and unbalanced label density. Therefore, the double annulus model and further division of the instances are used for DEPT, thus improving the performance.
3. As shown in Tables 2 and 3, for the higher-dimensional data sets in this experiment, such as medical, with 1449 dimensions, DEPT ranks first in more than 55% of cases in terms of 6 evaluation metrics, which benefits from the layered strategy. In terms of 7 numeric types of data sets, DEPT ranks first in more than 52% of cases. Furthermore, compared with LIFT, which constructed label-specific features by the k-means algorithm, DEPT performs better on more than 63% of experimental results. DEPT is superior to FRSLIFT in more than 60% of cases, which used sample selection based on LIFT's label-specific features. Compared with MULFE, which combined label-specific features and ensemble learning, DEPT performs better in more than 78% of cases. In addition, in terms of the average of 6 evaluation metrics, except for the result of one-error, which is close to LIFT, DEPT is superior in terms of macro-averaging AUC, average precision, coverage, ranking loss, and Hamming loss.
4. Figure 4 shows the spider web diagrams to intuitively illustrate the stability performance of each compared algorithm in terms of each evaluation metric. As shown in Figure 4, DEPT is more stable than the other algorithms. Namely, average precision and macro-averaging AUC tend to be larger, while coverage, ranking loss, Hamming loss, and one-error tend to be smaller.

To sum up, compared with other algorithms, under the condition that the AP is relatively stable, the additional evaluation metrics, such as CV, OE, and HL, have been improved to a certain extent, which finally verifies the effectiveness of the sample stratification and annulus clustering.

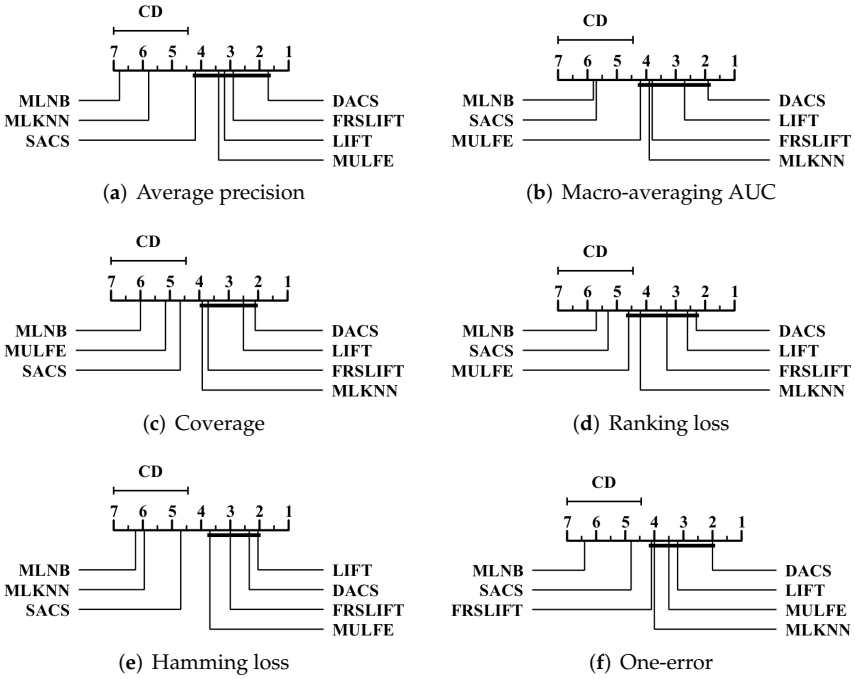


Figure 3. Comparison of DEPT (control algorithms) against other compared algorithms with the Bonferroni–Dunn test.

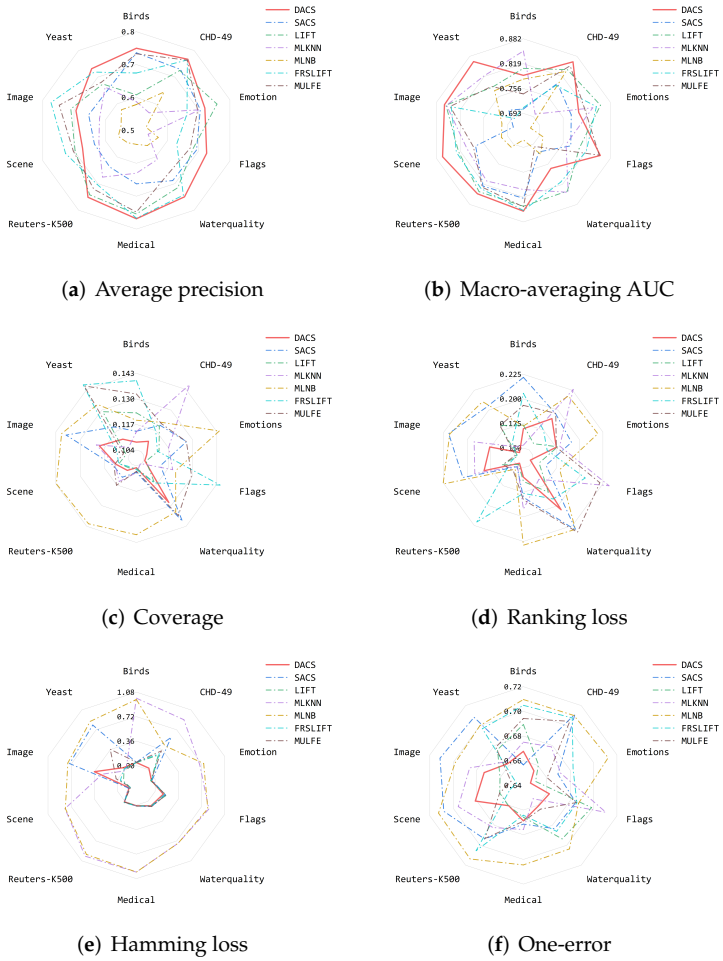


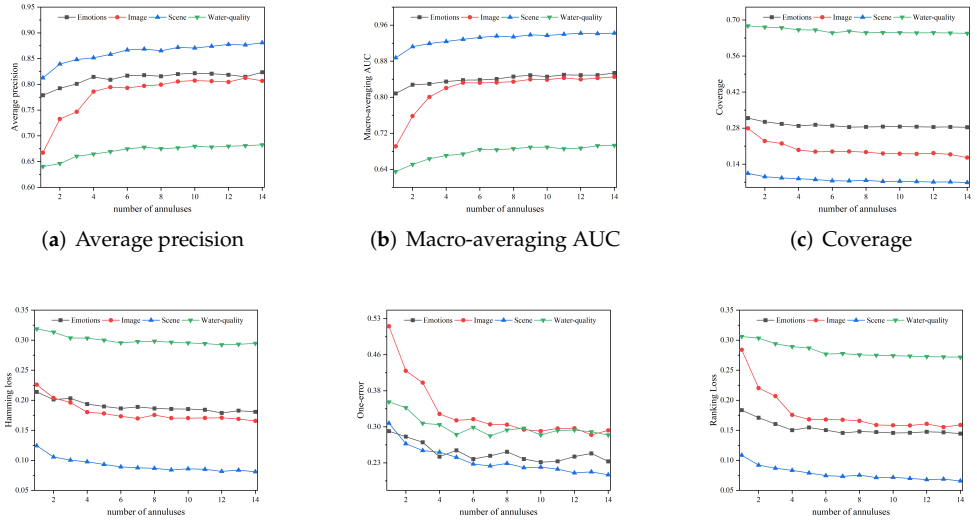
Figure 4. Spider web diagrams show the stability performance of compared algorithms with different evaluation metrics.

4.4. Further Analysis

4.4.1. Parameter Sensitivity

To further investigate the sensitivity of DEPT concerning the only parameter r, Figure 5 demonstrates the variation of each evaluation metric with the specified parameter, and the parameter value is sequentially set from 1 to 14. As shown in Figure 5, it is not difficult to draw the following conclusions:

1. In most cases, DEPT has relatively poor performance when the parameter is set to 1–2, mainly because the layering effect is hard to achieve on a small number of annuluses.
2. As the parameters increase sequentially, the performance improves from rapid increase to gradual stability.
3. The performance reaches its relative optimum as the parameter r increases beyond 12. Therefore, these conclusions justify the parameter setting of DEPT in the experimental parts.



(**a**) Average precision (**b**) Macro-averaging AUC (**c**) Coverage

(**d**) Hamming loss (**e**) One-error (**f**) Ranking loss

Figure 5. Performance of DEPT changes in terms of each evaluation metric as the parameter r increases from 1 to 14 on four regular-scale data set.

4.4.2. Execution Time

To study the runtime efficiency of SCFS and DEPT, Table 6 records the execution time of 5 algorithms with superior performance in Section 4.3. Combined with the experimental results, conclusions can be drawn as follows:

1. Since FRSLIFT conducted sample selection when constructing the feature space, a lot of time is sacrificed while improving performance.
2. Compared with LIFT, SCFS and DEPT can achieve shorter running time. This is mainly because the dimension of the constructed feature space φ_d is different, specifically, $\varphi_d(SCFS_k) = 2r, \varphi_d(DEPT_k) = 4r$, and $\varphi_d(LIFT_k) = \epsilon \cdot min(|POS|, |NEG|)$. Generally, the relationship can be written as follows:

$$\varphi_d(SCFS) < \varphi_d(DEPT) < \varphi_d(LIFT).$$

3. The label-specific feature space constructed based on annulus clustering can improve runtime efficiency. The difference is that the hierarchical structure of double annuluses is better than a single annulus in terms of 6 evaluation metrics, and time consumption is also more remarkable than that of the single annulus. In short, these results validate the efficiency of SCFS and DEPT in learning from multi-label data.

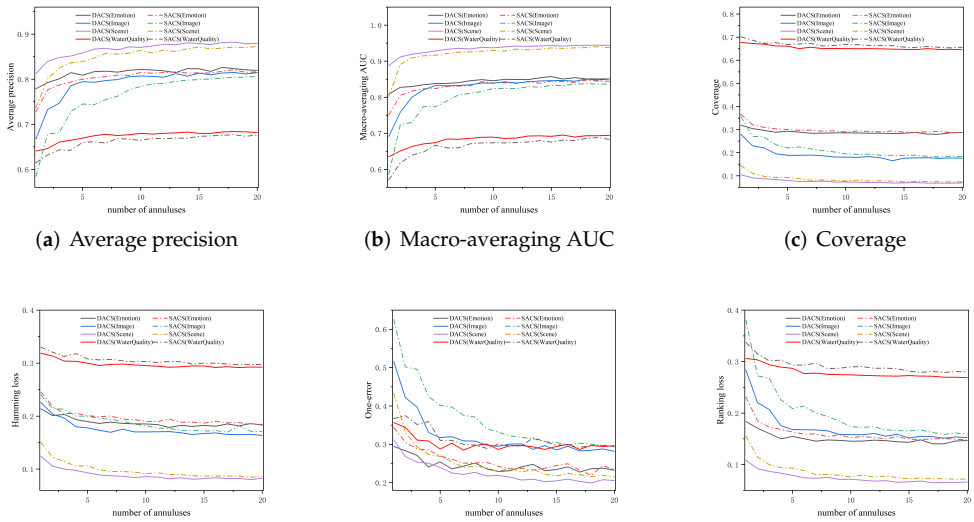
Table 6. Execution time of four compared algorithms (mean ± std) on 6 regular-scale data sets.

Execution Time	Data Set	SCFS	DEPT	LIFT	FRSLIFT
Total time (in seconds)	Emotions	0.227 ± 0.006	0.245 ± 0.010	0.301 ± 0.014	48.090 ± 0.157
	CHD49	0.373 ± 0.027	0.429 ± 0.036	0.744 ± 0.285	53.389 ± 0.781
	Images	2.672 ± 0.028	4.323 ± 0.057	6.287 ± 0.016	1879.930 ± 57.506
	WaterQuality	4.125 ± 0.142	5.436 ± 0.216	6.117 ± 0.106	727.960 ± 14.067
	Scene	5.082 ± 0.116	6.036 ± 0.144	6.175 ± 0.061	1599.054 ± 47.101
	Yeast	12.839 ± 0.156	24.689 ± 0.470	46.706 ± 0.736	17,198.588 ± 763.569

Formally, the time complexity of SCFS can be calculated from as follows. Firstly, the cost of calculating the distance between the instances and the center of annuluses is $O(l|PON_k|)$. The time spent dividing the instances in PON_k into annuluses is $O(lr)$. Then, the time spent performing clustering within annuluses is $lr(\mathcal{R}_k^*)^2$. Therefore, the total time complexity of SCFS is $O(l(|PON_k| + r + r(\mathcal{R}_k^*)^2))$. Analogously, the time complexity of DEPT is $O(l(t + |POS_k| + |NEG_k| + 2r + 2r(\mathcal{R}_k^*)^2))$, where t denotes the time of dimensional reduction of PCA [14] on POS_k and NEG_k. It is obvious that the time complexity of DEPT is higher than that of SCFS.

4.4.3. DEPT vs. SCFS

Both the DEPT and SCFS strategies can achieve the effect of a hierarchical cluster. To explore the difference in performance between the two strategies, we carried out comparative experiments on the two algorithms in terms of different evaluation metrics. Figure 6 reports the detailed experimental results. On the one hand, the two strategies conform to the monotonicity, i.e., either an upward or a downward trend on each evaluation metric. On the other hand, the performance gap between the two algorithms becomes more minor as the number of annuluses increases from 1 to 20.



(**a**) Average precision (**b**) Macro-averaging AUC (**c**) Coverage

(**d**) Hamming loss (**e**) One-error (**f**) Ranking loss

Figure 6. Comparison of DEPT and SCFS in terms of each evaluation metric as the parameter annulus increases from 1 to 20 on 4 benchmark multi-label data sets.

4.4.4. MCDA

Multi-criteria decision analysis (MCDA) [47] is an analytical method used to make choices among multiple criteria. In this paper, we use MCDA to solve conflicting requirements for speed and reliability of classification. Table 7 shows the calculated average accuracy rate, time consumption, and interaction index of all subsets in the annulus set and Mobius. In the fourth column of Table 7, Mobius corresponding to r_2, r_3, r_4, and r_5 will be used as a measure of information redundancy in the joint contribution of multiple attributes. The fifth column represents the Shapley value, which can be interpreted as a

measure of the importance of the corresponding attributes in the feature space, where r_1 and r_2 are complementary. The negative value of the corresponding influence index in this column indicates that the redundancy of the feature space increases with the increase in the number of annuluses. The fifth column can be similarly interpreted. From Table 7, it can be seen that the appropriate number of annuluses to balance time and efficiency is 5.

Table 7. Annuluses of DEPT in Mobius representation and mutual influence index.

Number of Annuluses	Time	Accuracy	Mobius	Shapley	Banzhaf
$\{r_1, r_2\}$	0.21	0.57	–0.67	0.67	0.67
$\{r_2, r_3\}$	0.34	0.62	0.17	0.32	0.29
$\{r_2, r_3, r_4\}$	0.66	0.66	0	0.11	0.11
$\{r_2, r_3, r_4, r_5\}$	1.35	0.75	0.32	0.44	0.5
$\{r_2, r_3, r_4, r_5, r_6\}$	2.76	0.77	0.67	–0.67	–0.67

5. Conclusions

A new method for extracting label-specific feature strategies, DEPT, is proposed in this paper. To form the double annulus domain, the centers of the double annulus field model are shifted to the sample center of the positive and negative samples, and the positive and negative samples are divided into annuluses, respectively. In addition, the principal component analysis (PCA) technique is utilized to divide the samples in each layer of annuluses. Based on this strategy, intra-annulus clustering was performed to further distinguish the two classes of instances close to the center of distribution density, aiming to improve the performance of multi-label classification with class-imbalanced. Comparative experiments with six algorithms on 10 multi-label data sets show the proposed strategy is superior to some others in 6 evaluation metrics. However, the multi-annulus model still faces the following problems and challenges:

1. Through the sensitivity analysis experiment in Section 4.4.1, we learned that increasing the number of annuluses can improve the overall classification accuracy but may lead to overfitting as the number of annuluses continues to increase.
2. Increasing the number of annuluses can improve classification accuracy to a certain extent, but it can also significantly impact efficiency because every additional annulus requires an exponential increase in training iterations by a factor of 2.
3. Based on the experimental results, the performance of the annulus model is not satisfactory for some data sets, which may be due to the single annulus model being unable to universally fit all data sets.

Future work will implement an annulus model that better fits various distribution trends of multi-label data sets and attempt to strike a balance between accuracy and efficiency.

Author Contributions: Conceptualization, Y.L. and C.L.; methodology, J.S.; software, Y.L.; validation, J.S., X.Y. and T.X.; formal analysis, P.W.; investigation, J.S.; resources, X.Y.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, J.S.; visualization, J.S.; supervision, X.Y.; project administration, X.Y.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 61906078, 62076111, 62006099) and the Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province (No. OBDMA202104).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, M.; Zhou, Z. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 1819–1837. [CrossRef]
2. Rubin, T.; Chambers, A.; Smyth, P.; Steyvers, M. Statistical topic models for multi-label document classification. *Mach. Learn.* **2012**, *88*, 157–208. [CrossRef]
3. Bromuri, S.; Zufferey, D.; Hennebert, J.; Schumacher, M. Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. *J. Biomed. Inform.* **2014**, *51*, 165–175. [CrossRef] [PubMed]
4. Trohidis, K.; Tsoumakas, G.; Kalliris, G.; Vlahavas, I. Multi-label classification of music by emotion. *EURASIP J. Audio Speech Music Process.* **2011**, *2011*, 4. [CrossRef]
5. Wu, B.; Zhong, E.; Horner, A.; Yang, Q. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 117–126.
6. Neville, J.; Jensen, D. Iterative classification in relational data. In Proceedings of the AAAI-2000 workshop Conference on Learning Statistical Models from Relational Data, Arlington, VA, USA, 12–15 May 2000; pp. 13–20.
7. Wu, T.; Fan, J.; Wang, P. An improved three-way clustering based on ensemble strategy. *Mathematics* **2022**, *10*, 1457. [CrossRef]
8. Wang, P.; Yao, Y. Ce3: A three-way clustering method based on mathematical morphology. *Knowl. Based Syst.* **2018**, *155*, 54–65. [CrossRef]
9. Zhang, M.; Wu, L. LIFT: Multi-label learning with label-specific features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 107–120. [CrossRef]
10. Huang, J.; Li, G.; Huang, Q.; Wu, X. Learning label specific features for multi-label classification. In Proceedings of the IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015; pp. 181–190.
11. Jia, X.; Zhu, S.; Li, W. Joint label-specific features and correlation information for multi-label learning. *J. Comput. Sci. Technol.* **2020**, *35*, 247–258. [CrossRef]
12. Xu, S.; Yang, X.; Yu, H.; Yu, D.; Yang, J.; Tsang, E. Multi-label learning with label-specific feature reduction. *Knowl.-Based Syst.* **2016**, *104*, 52–61. [CrossRef]
13. Niknam, T.; Amiri, B.; Olamaei, J.; Arefi, A. An efficient hybrid evolutionary optimization algorithm based on PSO and SA for clustering. *J. Zhejiang Univ. Sci. A* **2009**, *10*, 512–519. [CrossRef]
14. Abdi, H.; Williams, L. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
15. Zhang, P.; Gao, W.; Hu, J.; Li, Y. Multi-Label Feature Selection Based on High-Order Label Correlation Assumption. *Entropy* **2020**, *22*, 797. [CrossRef]
16. Zhang, Q.; Zhong, G.; Dong, J. A Graph-based Semi-supervised Multi-label Learning Method Based on Label Correlation Consistency. *Cogn. Comput.* **2021**, *13*, 1564–1573. [CrossRef]
17. Nguyen, H.; Woon, Y.; Ng, W. A survey on data stream clustering and classification. *Knowl. Inf. Syst.* **2015**, *45*, 535–569. [CrossRef]
18. Read, J.; Bifet, A.; Holmes, G.; Pfahringer, B. Scalable and efficient multi-label classification for evolving data streams. *Mach. Learn.* **2012**, *88*, 243–272. [CrossRef]
19. Braytee, A.; Liu, W.; Anaissi, A.; Kennedy, P. Correlated multi-label classification with incomplete label space and class imbalance. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–26. [CrossRef]
20. Liu, B.; Tsoumakas, G. Dealing with class imbalance in classifier chains via random undersampling. *Knowl. Based Syst.* **2020**, *192*, 105292. [CrossRef]
21. Fan, Y.; Liu, J.; Weng, W.; Chen, B.; Chen, Y.; Wu, S. Multi-label feature selection with local discriminant model and label correlations. *Neurocomputing* **2021**, *442*, 98–115. [CrossRef]
22. Liu, J.; Lin, Y.; Wu, S.; Wang, C. Online multi-label group feature selection. *Knowl. Based Syst.* **2018**, *143*, 42–57. [CrossRef]
23. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning multi-label scene classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [CrossRef]
24. Zhang, M.; Li, Y.; Liu, X.; Geng, X. Binary relevance for multi-label learning: An overview. *Front. Comput. Sci.* **2018**, *12*, 191–202. [CrossRef]
25. Elisseeff, A.; Weston, J. A kernel method for multi-labelled classification. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 681–687.
26. Mencía, E.; Furnkranz, J. Pairwise learning of multilabel classifications with perceptrons. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 2899–2906.
27. Brinker, C.; Mencía, E.; Fürnkranz, J. Graded multilabel classification by pairwise comparisons. In Proceedings of the 2014 IEEE International Conference on Data Mining, Shenzhen, China, 14–17 December 2014; pp. 731–736.
28. Yazici, V.O.; Gonzalez-Garcia, A.; Ramisa, A.; Twardowski, B.; Weijer, J.V.D. Orderless recurrent models for multi-label classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13440–13449.
29. Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier chains for multi-label classification. *Mach. Learn.* **2011**, *85*, 333–359. [CrossRef]
30. Tsoumakas, G.; Katakis, I.; Vlahavas, I. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 1079–1089. [CrossRef]
31. Song, J.; Tsang, E.; Chen, D.; Yang, X. Minimal decision cost reduct in fuzzy decision-theoretic rough set model. *Knowl. Based Syst.* **2017**, *126*, 104–112. [CrossRef]

32. Zhan, W.; Zhang, M. Multi-label learning with label-specific features via clustering ensemble. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21 October 2017; pp. 129–136.
33. Hang, J.; Zhang, M. Collaborative Learning of Label Semantics and Deep Label-Specific Features for Multi-Label Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 9860–9871. [CrossRef]
34. Che, X.; Chen, D.; Mi, J. A novel approach for learning label correlation with application to feature selection of multi-label data. *Inf. Sci.* **2020**, *512*, 795–812. [CrossRef]
35. Pei, G.; Wang, Y.; Cheng, Y.; Zhang, L. Joint label-density-margin space and extreme elastic net for label-specific features. *IEEE Access* **2019**, *7*, 112304–112317. [CrossRef]
36. Lin, Y.; Hu, Q.; Liu, J.; Zhu, X.; Wu, X. MULFE: Multi-label learning via label-specific feature space ensemble. *ACM Trans. Knowl. Discov. Data* **2021**, *16*, 1–24. [CrossRef]
37. Zhang, M.; Fang, J.; Wang, Y. BiLabel-Specific Features for Multi-Label Classification. *ACM Trans. Knowl. Discov. Data* **2021**, *16*, 1–23. [CrossRef]
38. Godbole, S.; Sarawagi, S. Discriminative methods for multi-labeled classification. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 26–28 May 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 22–30.
39. Schapire, R.; Singer, Y. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* **1999**, *37*, 297–336. [CrossRef]
40. Zhang, M.; Zhou, Z. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
41. Schapire, R.; Singer, Y. BoosTexter: A boosting-based system for text categorization. *Mach. Learn.* **2000**, *39*, 135–168. [CrossRef]
42. Salton, G. Developments in automatic text retrieval. *Science* **1991**, *253*, 974–980. [CrossRef]
43. Zhang, M.; Peña, J.; Robles, V. Feature selection for multi-label naive Bayes classification. *Inf. Sci.* **2009**, *179*, 3218–3229. [CrossRef]
44. Chang, C.; Lin, C. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
45. Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* **1940**, *11*, 86–92. [CrossRef]
46. Dunn, O. Multiple comparisons among means. *J. Am. Stat. Assoc.* **1961**, *56*, 52–64. [CrossRef]
47. Montibeller, G.; Franco, A. Multi-criteria decision analysis for strategic decision making. In *Handbook of Multicriteria Analysis*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 25–48.