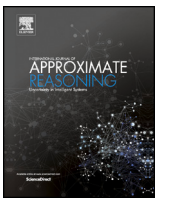
International Journal of Approximate Reasoning 169 (2024) 109181

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar



Feature selection for multi-label learning based on variable-degree multi-granulation decision-theoretic rough sets

Ying Yu [a,b,*], Ming Wan [b], Jin Qian [b], Duoqian Miao [c], Zhiqiang Zhang [b], Pengfei Zhao [d]

[a] State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure, East China Jiaotong University, Nanchang, 330013, Jiangxi, China
[b] College of Software, East China Jiaotong University, Nanchang, 330013, Jiangxi, China
[c] School of Electronic and Information Engineering, Tongji University, Shanghai, 210048, China
[d] Digital Economy Research Institute of Jiangxi Provincial Investment Group, Nanchang, 330013, Jiangxi, China

ARTICLE INFO

Keywords:
Three-way decision
Multi-label learning
Feature selection
Rough sets
Uncertainty

ABSTRACT

Multi-label learning (MLL) suffers from the high-dimensional feature space teeming with irrelevant and redundant features. To tackle this, several multi-label feature selection (MLFS) algorithms have emerged as vital preprocessing steps. Nonetheless, existing MLFS methods have their shortcomings. Primarily, while they excel at harnessing label-feature relationships, they often struggle to leverage inter-feature information effectively. Secondly, numerous MLFS approaches overlook the uncertainty in the boundary domain, despite its critical role in identifying high-quality features. To address these issues, this paper introduces a novel MLFS algorithm, named VMFS. It innovatively integrates multi-granulation rough sets with three-way decision, leveraging multi-granularity decision-theoretic rough sets (MGDRS) with variable degrees for optimal performance. Initially, we construct coarse decision (RDC), fine decision (RDF), and uncertainty decision (RDU) functions for each object based on MGDRS with variable degrees. These decision functions then quantify the dependence of attribute subsets, considering both deterministic and uncertain aspects. Finally, we employ the dependency to assess attribute importance and rank them accordingly. Our proposed method has undergone rigorous evaluation on various standard multi-label datasets, demonstrating its superiority. Experimental results consistently show that VMFS significantly outperforms other algorithms on most datasets, underscoring its effectiveness and reliability in multi-label learning tasks.

1. Introduction

In the traditional machine learning tasks, it is often assumed that each object is given only one category label, which is referred to as single-label learning. However, in reality an object may be associated with multiple labels. For example, a patient may be related to multiple diseases in medical diagnosis. In video classification, a video may belong to multiple categories, such as horror, action and romance. In automatic image annotation, an image can simultaneously have multiple labels, such as blue sky, white clouds and

* Corresponding author.
E-mail address: yuyingjx@163.com (Y. Yu).

https://doi.org/10.1016/j.ijar.2024.109181
Received 22 August 2023; Received in revised form 29 February 2024; Accepted 19 March 2024

Available online 27 March 2024

Fig. 1. Images with one object.

green water. Traditional methods can no longer be employed to deal with such multi-semantic objects, which led to the emergence of numerous multi-label learning (MLL) algorithms [1–4].

Similar to traditional single-label learning, multi-label learning also suffers from the curse of dimensionality, which not only increases the computational complexity, but also increases the difficulty of modelling and decision making. Additionally, the high-dimensional feature space of multi-label learning also introduces a significant amount of redundant and noisy information, which can adversely impact the performance of classifiers. Consequently, in order to improve the efficiency of the multi-label algorithm and reduce the impact of redundancy and noise, it is necessary to propose effective methods that can reduce the dimensionality of high-dimensional multi-label data.

One way to deal with the high-dimensional feature problems is multi-label feature extraction, the main idea of which is to transform the original high-dimensional feature space into a lower-dimensional feature space that preserves as much information as possible from the original features. For example, the MDDM algorithm proposed by Zhang [5] utilizes the mapping space and subspace for dimensionality reduction, using two strategies, linear kernel and nonlinear kernel, respectively. The LDA algorithm proposed by Sun [6] utilizes the idea of problem transformation to convert a multi-label problem into a single-label problem and then uses the single-label method for feature reduction. However, the LDA algorithm does not take into account the relationship or interdependence between the labels.

Another way to deal with high-dimensional feature spaces is multi-label feature selection (MLFS) [7,8]. Compared with feature extraction, feature selection can retain the physical meaning of the features themselves and has perfect interpretability, so it has received more extensive attention from researchers. In the field of machine learning, feature selection methods are commonly categorized into three types: wrapper [9,10], filter [11,12], and embedded [13] algorithms. As a subfield of machine learning, multi-label learning similarly adopts these methods to classify multi-label feature selection algorithms. Among them, filter algorithms are the most commonly used for the multi-label feature selection due to their relative independence from specific classifiers, their low computational effort, and their high generalization capability. The commonly used measures for filter strategies include distance measure, dependency measure, mutual information measure, and consistency measure.

Due to the multi-semantic nature of multi-label objects, there is a significant amount of uncertainty or ambiguity in multi-label data. Taking Fig. 1 as an example, these two images containing only one object, sharing similar visual features such as texture and contour, yet their semantic labels are completely different. The image on the right is labelled as “woman”, while the one on the left is labelled as “dog”. Due to the similarity in features, it is easy to confuse them during recognition. Such images with only one label are typically referred to as single-label images, which can be seen as a special case of multi-label images. Since single-label images generally contain only one object, the recognition is simpler compared to multi-label images. As the number of objects in the images increases, the number of semantic labels also increase, making the problem more complex. Furthermore, as illustrated in Fig. 2, the scale of objects in the image varies significantly. For instance, individuals positioned far from the camera or small-sized backpacks may be prone to misidentification due to limited pixels or insufficient clarity within the image, leading to significant uncertainty in the recognition results. In a medical diagnosis, a doctor would base his or her diagnosis on the patient’s symptoms. If a patient has vomiting, dizziness, or other symptoms, the doctor may decide that the patient has a cold, heat stroke, high blood pressure, or a combination of these. The uncertainty in diagnosis primarily stems from incomplete information regarding the symptoms. It is evident that uncertainty or ambiguity is prevalent in multi-label data.

Feature selection has gained wider recognition for its unique advantages, which has prompted scholars to design multi-label feature selection algorithms from different perspectives. Zhang et al. [14] introduced MLNB, seeking the optimal features by using principal component analysis and genetic algorithm. Lin et al. [15] proposed a norm regularization-based method, which considers label correlations and achieves embedded multi-label feature selection through low-dimensional compression. Fan et al. [16] presented LCIFS, utilizing a manifold framework and a regression model to fit feature space-label distribution relationships. Although the above algorithms have achieved some success, there are still great challenges in analyzing uncertainties such as ambiguities or inconsistencies in multi-label data.

Fig. 2. Images with diverse objects.

Rough sets theory [17] is an effective tool for dealing with uncertainty problems, which does not require any prior information other than the data. It was proposed by Polish scientist Pawlak and is widely applied in single-label feature selection algorithms [18, 19]. Recently, many researchers have also extended rough set theory for MLFS [20–22]. For example, Duan [23] used neighbourhood rough sets to build upper and lower neighbourhood approximations and proposed an MLFS algorithm based on neighbourhood rough sets (MNRS). Lin [24] used neighbourhood rough sets to calculate the neighbourhood mutual information between labels and features, and proposed an MLFS algorithm based on neighbourhood mutual information which considers uncertainty. Li [25] proposed an MLFS algorithm based on variable precision rough sets, which can accurately catch the implied uncertainty associated with labels. Liang [26] proposed an MLFS algorithm based on optimal granulation selection, which fully considers the uncertainty implied by labels. However, in multi-label data there are correlations not only between features and labels, but also between labels. Therefore, inter-label correlation needs to be taken into account for feature selection. Xu [27] proposed an MLFS algorithm (MFSFN) based on fuzzy neighbourhood rough sets by combining fuzzy sets and neighbourhood rough sets and proposed a hybrid measurement tactic by combining fuzzy neighbourhood conditional entropy and fuzzy neighbourhood approximation accuracy. The multi-label feature selection algorithm LDRS proposed by Liu [21] integrates label distribution and neighbourhood rough sets to solve the problem of significant label differences among instances, achieving excellent performance on both public and real-world datasets. Although these rough set-based multi-label feature selection algorithms have achieved some improvements in considering correlation and uncertainty, most of them measure the importance of features in a single granularity space. However, there may exist complex correlations between features and labels, features and features, or labels and labels in multi-label data, which require a comprehensive description through multi-granularity spaces.

In the view of granular computing, the classical rough set theory is established through a single granulation, where the upper/lower approximations of the target concepts are approximated via single relations on the universe. However, in some practical situations, it is necessary to describe a target concept simultaneously through multiple binary relations on the universe. In order to apply rough set theory more widely to practical issues, Qian [28] proposed multi-granulation decision-theoretic rough sets (MGDRS) which incorporates three-way decision theory [29] into multi-granulation rough sets and solves the problem by transforming a large feature space into multiple smaller granular spaces through a partition-like operation. MGDRS is an effective data modelling theory based on multiple granularity spaces. Similar to weakly supervised learning, and especially to superset learning [30,31], it defines both optimistic and pessimistic models. The “optimistic” one interprets uncertain data in a way that is most favorable for candidate models, resulting in loosely defined upper and lower approximations in the optimistic MGDRS model. Conversely, the “pessimistic” one guides model selection with the least favorable interpretation, leading to overly strict upper and lower approximations in the pessimistic MGDRS model. Both models exhibit extreme tendencies and struggle to adapt flexibly to the requirements of multi-label feature selection.

To enhance the flexibility of MGDRS for MLFS tasks, we introduce a variable degree to the MGDRS model, aiming to strike a balance between the pessimistic and optimistic models. Consequently, we propose a variable-degree MGDRS-based MLFS algorithm (VMFS) falling under the filter type. The proposed algorithm integrates the concepts of multi-granulation rough sets and three-way decision, thereby rendering the feature selection process more adaptable and interpretable. In summary, the article’s primary contributions can be outlined as follows:

- The fine decision function, the uncertain decision function and the coarse decision function are defined based on the multi-granulation decision-theoretic rough sets with the variable degree to fully analyze the certain and uncertain information of the data.
- The variable degree parameter υ between pessimistic and optimistic degree of MDGRS was dynamically adjusted to accommodate different types of multi-label datasets.
-To quantify the dependency and uncertainty of features on labels for feature selection purposes, we have developed an label approximate accuracy function by leveraging the coarse decision function and the fine decision function. During the construction of the label approximate accuracy function, we adopt a multi-granularity approach to fully exploit the information encapsulated within the features and analyze the correlation information among them.
-The proposed MLFS algorithm based on the variable degree MGDRS achieves good results on most datasets.

2. Preliminaries

2.1. Three-way decision

The three-way decision [29] is a decision-making model that mimics human cognition. In the process of decision-making, people make immediate judgements of acceptance or rejection for things that they are sufficiently sure of, while they tend to use delayed decision-making for things that they cannot immediately judge. Unlike the traditional binary decision-making framework, three-way decision provides three decision outcomes with higher degrees of freedom and fault tolerance.

In a straightforward knowledge representation scheme, a finite set of objects is characterized by a finite set of attributes. This scheme can be formally defined using an information table S, expressed as a tuple.

$$S = \left(U, At, \left\{V_a | a \in At\right\}, \left\{I_a | a \in At\right\}\right), \tag{1}$$

where U is a finite nonempty set of objects, At is a finite nonempty set of attributes, V_a is a nonempty set of values for an attribute $a \in At$, and $I_a : U \to V_a$ is an information or description function. It is assumed that the mapping I_a is single-valued. In this scenario, the value of an object $x \in U$ for an attribute $a \in At$ is represented by $I_a(x)$.

For any subset of attributes $A \subseteq At$, an equivalence relation $Ind(A)$ on U can be defined as follows:

$$x_1 \ Ind(A) \ x_2 \Longleftrightarrow \forall a \in A[I_a(x_1) = I_a(x_2)]. \tag{2}$$

In other words, two objects x_1 and x_2 possess identical attribute values for every attribute defined in A, then they are characterized as being indistinguishable with respect to A. The equivalence class containing object x is denoted $[x]_A$ or $[x]$.

The attribute set can be divided into two subsets, namely the set of condition attributes C and the set of decision attributes D, denoted as $At = C \cup D$, where C and D are non-overlapping subsets and $C \cap D = \emptyset$. For simplicity, Let $\pi_C = \{c_1, c_2, \ldots, c_m\}$ represent m disjoint condition classes defined by the condition attribute set C, and $\pi_D = \{d_1, d_2, \ldots d_n\}$ represent n disjoint decision classes defined by the decision attribute set D.

In the Pawlak[13] approximation space, the object x is usually represented by the equivalence class $[x]$. Since inconsistent data often occur in practice, an object $[x]_A$ is usually classified as correctly as possible into the positive region $POS(X)$, the boundary region $BND(X)$, and the negative region $NEG(X)$, where $X \subseteq U$. Based on Bayesian theory and the minimum risk criterion, there is a special case where the loss function has to meet certain conditions $\xi_{pp} \leqslant \xi_{bp} < \xi_{np}$ and $\xi_{nn} \leqslant \xi_{bn} < \xi_{pn}$. Thus, two thresholds and can be calculated ($0 \leqslant \beta < \alpha \leqslant 1$), namely:

$$\alpha = \frac{\left(\xi_{pn} - \xi_{bn}\right)}{\left(\xi_{pn} - \xi_{bn}\right) + \left(\xi_{bp} - \xi_{pp}\right)}, \beta = \frac{\left(\xi_{bn} - \xi_{nn}\right)}{\left(\xi_{bn} - \xi_{nn}\right) + \left(\xi_{np} - \xi_{bp}\right)}, \tag{3}$$

where $\xi_{pp}, \xi_{bp}, \xi_{np}, \xi_{nn}, \xi_{bn}, \xi_{pn}$ denote the losses incurred for taking actions P, B and N when x belong to X or not. In the decision table S, for a decision class $d_j \in \pi_D$, the probabilistic lower approximation set and probabilistic upper approximation set relative to are defined as follows:

$$\underline{C^{(\alpha,\beta)}(d_j)} = \bigcup\nolimits_{x \in U} \{x | \Pr(d_j | [x]_C) \geqslant \alpha\}, \tag{4}$$

$$\overline{C^{(\alpha,\beta)}(d_j)} = \bigcup\nolimits_{x \in U} \{x | \Pr(d_j | [x]_C) > \beta\}, \tag{5}$$

$$\Pr\left(d_j | [x]_A\right) = \frac{\left|d_j \cap [x]_C\right|}{\left|[x]_C\right|}, \tag{6}$$

where $\Pr\left(d_j | [x]_C\right)$ denotes the conditional probability that an object x belonging to a decision class d_j. Thus in decision rough set model π_D, the three probability regions can be expressed as:

$$POS^{(\alpha,\beta)}\left(\pi_D \mid \pi_C\right) = \bigcup\nolimits_{x \in U} \{x | \Pr\left(d_{\max}\left([x]_C\right) | [x]_C\right) \geqslant \alpha\}, \tag{7}$$

$$NEG^{(\alpha,\beta)}\left(\pi_D \mid \pi_C\right) = \bigcup\nolimits_{x \in U} \{x | \Pr\left(d_{\max}\left([x]_C\right) | [x]_C\right) \leqslant \beta\}, \tag{8}$$

$$BND^{(\alpha,\beta)}\left(\pi_D \mid \pi_C\right) = \bigcup\nolimits_{x \in U} \left\{x | \beta < \Pr\left(d_{\max}\left([x]_C\right) | [x]_C\right) < \alpha\right\}, \tag{9}$$

where $d_{\max}\left([x]_C\right) = \arg\max_{d_j \in \pi_D} \{\frac{\left|[x]_C \cap d_j\right|}{|[x]_C|}\}$ denotes the most dominant decision class in the equivalence class $[x]_C$.

The three-way decision has also been widely applied to multi-label learning tasks. For instance, Zhang [32] proposed a multi-label classification algorithm based on granular structure using the concept of three-way decision and obtained better classification results on several multi-label datasets. Zhao [33] proposed an intuitionistic fuzzy set-based label enhancement algorithm for multi-label classification by combining three-way decision-making and fuzzy sets, which performs well in terms of accuracy.

Table 1
Multi-label decision table.

	C_1	C_2	C_3	C_4	l_1	l_2	l_3
x_1	1	2	1	2	1	0	1
x_2	2	2	2	3	0	1	0
x_3	1	2	1	2	1	0	1
x_4	2	3	1	3	1	1	1
x_5	2	3	1	3	0	0	1
x_6	1	2	2	3	0	0	0
x_7	2	3	1	3	1	1	1
x_8	1	2	2	1	1	1	0
x_9	1	2	2	1	0	0	0
x_{10}	3	2	2	3	1	1	1
x_{11}	1	1	2	2	1	1	0

2.2. Multi-granulation decision-theoretic rough sets

Suppose there is a decision table S, $C = \{C_1, C_2, \cdots, C_m\}$ is a set of attribute subsets. Then the lower and upper approximations of the optimistic MGRDS model of decision class $d_j \in \pi_D$ with the attribute subsets $C_1, C_2, \cdots, C_m$ can be defined as follows:

$$\overline{\sum_{i=1}^{m} C_i}^{o,(\alpha,\beta)}(d_j) = \bigcup_{x\in U}\left\{x|\Pr\left(d_j|[x]_{C_1}\right) \geqslant \alpha \cup \cdots \Pr\left(d_j|[x]_{C_m}\right) \geqslant \alpha\right\}, \tag{10}$$

$$\begin{aligned}\overline{\sum_{i=1}^{m} C_i^{o,(\alpha,\beta)}}(d_j) &= U - \bigcup_{x\in U}\left\{x \mid \Pr\left(d_j \mid [x]_{C_1}\right) \leqslant \beta \cap \cdots \Pr\left(d_j \mid [x]_{C_m}\right) \leqslant \beta\right\} \\ &= \bigcup_{x\in U}\{x \mid \Pr\left(d_j \mid [x]_{C_1}\right) > \beta \cup \cdots \Pr\left(d_j \mid [x]_{C_m}\right) > \beta,\end{aligned} \tag{11}$$

where α and β is two thresholds of three-way decision. Then the lower and upper approximations of the pessimistic MGRDS model of decision class $d_j \in \pi_D$ with the attribute subsets $C_1, C_2, \cdots, C_m$ can be defined as follows:

$$\overline{\sum_{i=1}^{m} C_i}^{p,(\alpha,\beta)}(d_j) = \bigcup_{x\in U}\left\{x \mid \Pr\left(d_j|[x]_{C_1}\right) \geqslant \alpha \cap \cdots \Pr\left(d_j|[x]_{C_m}\right) \geqslant \alpha\right\}. \tag{12}$$

$$\begin{aligned}\overline{\sum_{i=1}^{m} C_i^{p,(\alpha,\beta)}}(d_j) &= U - \bigcup_{x\in U}\left\{x|\Pr\left(d_j|[x]_{C_1}\right) \geqslant \beta \cup \cdots \Pr\left(d_j|[x]_{C_m}\right) \geqslant \beta\right\} \\ &= \bigcup_{x\in U}\left\{x \mid \Pr\left(d_j \mid [x]_{C_1}\right) > \beta \cap \cdots \Pr\left(d_j \mid [x]_{C_m}\right) > \beta\right\}.\end{aligned} \tag{13}$$

3. Proposed approaches

In this section, we first analyze the shortcomings of traditional pessimistic and optimistic multi-granulation decision-theoretic rough sets in constructing multi-label decision systems. Then, we delve into the construction of a multi-label decision system using variable-degree MGDRS, and provide a detailed description of implementing multi-label feature selection based on variable-degree MGDRS.

3.1. Description of notations

The multi-label information system can be denoted by a multi-label decision table $S = (U, A = C \cup L, V)$, where universe of discourse $U = \{x_1, x_2, ...x_n\}$ is a set of objects, $C = \{c_1, c_2, ...c_m\}$ represents a feature set and the label set is $L = \{l_1, l_2, \ldots l_q\}$. For any $l \in L, l$ takes a value between 0 and 1. For any attribute $a \in A$, whether it is a conditional attribute or a decision attribute, it is assumed that the set of all possible values corresponding to a is V_a. Then, V represents the union of all V_a, denoted by $V = \bigcup_{a\in A}\{V_a\}$.

To simplify the problem, a simple multi-label data decision table is constructed as shown in Table 1, containing a total of 11 objects, with each object having four attributes and three category labels.

3.2. Multi-label decision function based on pessimistic and optimistic decision-theoretic rough sets

Define 1. Let $S = (U, C \cup L, V)$ be a multi-label decision table, $C = \{c_1, c_2, ...c_m\}$, and $L = \{l_1, l_2, ..., l_q\}$. For any $l_i \in L$, an object belonging to the set of class l_i is defined as follows:

$$E_i = \bigcup_{x \in U} \left\{ x \mid l_i(x) = 1 \right\}, l_i \in L, \tag{14}$$

$$E = \{E_1, E_2, \ldots, E_q\}, \tag{15}$$

where $l_i(x) = 1$ represents x having l_i and $l_i(x) = 0$ means having no label l_i. Since each label is associated with at least one object in MLL, the following relationship can be proven:

$$\sum_{i=1}^{q} E_i = U. \tag{16}$$

Define 2. Let $S = (U, C \cup L, V)$ be a multi-label decision table, $B = \{B_1, B_2, \ldots, B_j\}$ is a set of j granules, $B_k \subseteq C$, $k \in \{1, ..., j\}$. The fine decision function $RDF^{\circ}_{\{\alpha,\beta\}}(x)$, the coarse decision function $RDC^{o}_{(\alpha,\beta)}(x)$ and the uncertainty decision function $RDU^{\circ}_{(\alpha,\beta)}(x)$ based on the optimistic MGDRS[27] are given as follows:

$$RDF^{\circ}_{\{\alpha,\beta\}}(x) = \bigcup_{i=1}^{q} \bigcup_{B_k \in B} \{l_i \mid \Pr(E_i \mid [x]_{B_k}) \geq \alpha\} = \bigcup_{i=1}^{q} \{l_i \mid \Pr(E_i \mid [x]_{B_1}) \geq \alpha \cup \ldots \cdots \Pr(E_i \mid [x]_{B_k}) \geq \alpha\}, \tag{17}$$

$$RDC^{o}_{(\alpha,\beta)}(x) = \bigcup_{i=1}^{q} \bigcup_{B_k \in B} \{l_i \mid \Pr(E_i \mid [x]_{B_k}) > \beta = \bigcup_{i=1}^{q} \{l_i \mid \Pr(E_i \mid [x]_{B_1}) > \beta \cup \ldots \Pr(E_i \mid [x]_{B_k}) > \beta\}, \tag{18}$$

$$RDU^{\circ}_{(\alpha,\beta)}(x) = RDF^{\circ}_{(\alpha,\beta)} - RDC^{\circ}_{(\alpha,\beta)}, \tag{19}$$

$$\Pr(E_i \mid [x]_{B_k}) = \frac{\left\| E_i \cap [x]_{B_k} \right\|}{\left\| [x]_{B_k} \right\|}, \tag{20}$$

where α and β are two thresholds in the three-way decision that satisfies $\alpha > \beta$. $\Pr(E_i \mid [x]_{B_k})$ represents the conditional probability of E_i under $[x]_{B_k}$ conditions. For each label l_i in the collection that generated by the coarse decision function RDC, there exists at least one granule B_k such that the conditional probability $\Pr(E_i \mid [x]_{B_k})$ is greater than β. For each label l_i in the collection that generated by the fine decision function RDF, there exists at least one granules B_k such that the conditional probability $\Pr(E_i \mid [x]_{B_k})$ is greater than or equal to α. The labels generated by the coarse decision function RDC are considered to be potentially related to the object x, while the labels generated by the fine decision function RDF are considered to be closely related to the object x. Those labels obtained by the uncertainty decision function RDU cannot be determined whether it is related to the object x or not.

Define 3. Let $S = (U, C \cup L, V)$ be a multi-label decision table, $B = \{B_1, B_2, ..., B_j\}$ is a set of j granules, $B_k \subseteq C$, $k \in \{1, ..., j\}$. The fine decision function $RDF^{p}_{(\alpha,\beta)}(x)$, the coarse decision function $RDC^{p}_{(\alpha,\beta)}(x)$ and the uncertainty decision function $RDU^{p}_{(\alpha,\beta)}(x)$ based on the pessimistic MGDRS[27] are given as follows:

$$RDF^{p}_{(\alpha,\beta)}(x) = \bigcup_{i=1}^{q} \bigcap_{B_k \in B} \{l_i \mid \Pr(E_i \mid [x]_{B_k}) \geq \alpha\} = \bigcup_{i=1}^{q} \{l_i \mid \Pr(E_i \mid [x]_{B_1}) \geq \alpha \cap \ldots \Pr(E_i \mid [x]_{B_k}) \geq \alpha\}, \tag{21}$$

$$RDC^{p}_{(\alpha,\beta)}(x) = \bigcup_{i=1}^{q} \bigcap_{B_k \in B} \{l_i \mid \Pr(E_i \mid [x]_{B_k}) > \beta\} = \bigcup_{i=1}^{q} \{l_i \mid \Pr(E_i \mid [x]_{B_1}) > \beta \cap .. \Pr(E_i \mid [x]_{B_k}) > \beta\}\ , \tag{22}$$

$$RDU^{p}_{(\alpha,\beta)}(x) = RDF^{p}_{(\alpha,\beta)} - RDC^{p}_{(\alpha,\beta)}, \tag{23}$$

$$\Pr(E_i \mid [x]_{B_k}) = \frac{\left\| E_i \cap [x]_{B_k} \right\|}{\left\| [x]_{B_k} \right\|}. \tag{24}$$

For each label l_i in the collection that generated by the coarse decision function RDC, the conditional probability $\Pr(E_i \mid [x]_{B_k})$ corresponding to all granules B_k must be greater than β. For each label l_i in the collection that generated by the fine decision function RDF, the conditional probability $\Pr(E_i \mid [x]_{B_k})$ corresponding to all granules B_k must be greater than α. Similar to the optimistic multi-granulation decision, the labels generated by the coarse decision function RDC are considered to maybe have correlation with the object x. The labels generated by the fine decision function RDF are considered to have a strong correlation with object x, while the labels generated by the uncertainty decision function RDU function cannot determine whether it is related to the object x or not.

Example 1. According to the multi-label decision table $S = (U, C \cup L, V)$ given in Table 1, the RDF function and RDC function based on optimistic condition and pessimistic condition are calculated as Table 2 shown, where $\alpha = 0.6, \beta = 0.4$.

From the above calculation results, it can be seen that under the optimistic condition, the RDC function and the RDF function generate many values, the decision boundary is not obvious, and the uncertainty cannot be well measured. Under the pessimistic condition, the RDC functions contain so few values, and some RDF functions even generate null values. Thus, the optimistic

Table 2
The calculation result for Example 1.

Optimistic condition	Pessimistic condition
$RDC^{o}_{(\alpha,\beta)}(x_1) = RDC^{o}(x_3) = \{l_1, l_2, l_3\}$	$RDC^{p}_{(\alpha,\beta)}(x_1) = RDC^{p}_{(\alpha,\beta)}(x_3) = \{l_1\}$
$RDC^{o}_{(\alpha,\beta)}(x_2) = RDC^{o}(x_{10}) = \{l_1, l_2, l_3\}$	$RDC^{p}_{(\alpha,\beta)}(x_2) = RDC^{p}_{(\alpha,\beta)}(x_{10}) = \{l_1, l_2\}$
$RDC^{o}_{(\alpha,\beta)}(x_4) = RDC^{o}(x_5) = RDC^{o}(x_7) = \{l_1, l_2, l_3\}$	$RDC^{p}_{(\alpha,\beta)}(x_4) = RDC^{p}_{(\alpha,\beta)}(x_5) = RDC^{p}_{(\alpha,\beta)}(x_7) = \{l_1, l_2, l_3\}$
$RDC^{o}_{(\alpha,\beta)}(x_6) = \{l_1, l_2, l_3\}$	$RDC^{p}_{(\alpha,\beta)}(x_6) = \{l_1\}$
$RDC^{o}_{(\alpha,\beta)}(x_8) = \{l_1, l_2, l_3\}$	$RDC^{p}_{(\alpha,\beta)}(x_8) = \{l_1\}$
$RDC^{o}_{(\alpha,\beta)}(x_9) = \{l_1, l_2\}$	$RDC^{p}_{(\alpha,\beta)}(x_9) = \{l_1\}$
$RDC^{o}_{(\alpha,\beta)}(x_{11}) = \{l_1, l_2, l_3\}$	$RDC^{p}_{(\alpha,\beta)}(x_{11}) = \{l_1\}$
$RDF^{o}_{(\alpha,\beta)}(x_1) = RDF^{o}_{(\alpha,\beta)}(x_3) = \{l_1, l_3\}$	$RDF^{p}_{(\alpha,\beta)}(x_1) = RDF^{p}_{(\alpha,\beta)}(x_3) = \{l_1\}$
$RDF^{o}_{(\alpha,\beta)}(x_2) = RDF^{o}_{(\alpha,\beta)}(x_{10}) = \{l_1, l_2, l_3\}$	$RDF^{p}_{(\alpha,\beta)}(x_2) = RDF^{p}_{(\alpha,\beta)}(x_{10}) = \{\}$
$RDF^{o}_{(\alpha,\beta)}(x_4) = RDF^{o}_{(\alpha,\beta)}(x_5) = RDF^{o}_{(\alpha,\beta)}(x_7) = \{l_1, l_2, l_3\}$	$RDF^{p}_{(\alpha,\beta)}(x_4) = RDF^{p}_{(\alpha,\beta)}(x_5) = RDF^{p}_{(\alpha,\beta)}(x_7) = \{l_2\}$
$RDF^{o}_{(\alpha,\beta)}(x_6) = \{l_1, l_2\}$	$RDF^{p}(x_6) = \{\}$
$RDF^{o}_{(\alpha,\beta)}(x_8) = \{l_1, l_2\}$	$RDF^{p}_{(\alpha,\beta)}(x_8) = \{\}$
$RDF^{o}_{(\alpha,\beta)}(x_9) = \{l_1, l_2\}$	$RDF^{p}_{(\alpha,\beta)}(x_9) = \{\}$
$RDF^{o}_{(\alpha,\beta)}(x_{11}) = \{l_1, l_2\}$	$RDF^{p}_{(\alpha,\beta)}(x_{11}) = \{\}$

condition is too loose, while the pessimistic condition is too strict. So neither the optimistic MGDRS nor the pessimistic MGDRS can effectively characterize the objects in the multi-label decision table accurately.

3.3. Multi-label decision function based on variable-degree multi-granulation decision-theoretic rough sets

Based on the previous section's discussion, it can be inferred that the conventional approaches of creating decision system via pessimistic and optimistic MGDRS have certain limitations. To better construct a multi-label decision system, a variable-degree MGDRS is proposed. It enables the decision stringency to vary between pessimistic and optimistic, which can be better adapted to different multi-label data.

Define 4. Let $S = (U, C \cup L, V)$ be a multi-label decision table, $B = \{B_1, B_2, \ldots, B_j\}$ is a set of j granules, $B_k \subseteq C$, $k \in \{1, ..., j\}$. The coarse decision operator r_i^1 and the fine decision operator r_i^2 are defined as follows:

$$r_i^1(x) = \bigcup_{k=1}^{j} \{B_k \mid \Pr(E_i \mid [x]_{B_k}) > \beta\}, E_i \in E, \tag{25}$$

$$r_i^2(x) = \bigcup_{k=1}^{j} \left\{B_k \mid \Pr(E_i \mid [x]_{B_k}) \geq \alpha\right\}, E_i \in E, \tag{26}$$

$$Pr(E_i \mid [x]_{B_k}) = \frac{E_i \bigcap [x]_{B_k}}{[x]_{B_k}}, \tag{27}$$

where α and β represents two thresholds in the three-way decision. r_i^1 represents the set of all granules B_k that satisfy $\Pr(E_i \mid [x]_{B_k}) > \beta$. r_i^2 represents the set of all granules B_k that satisfy $\Pr(E_i \mid [x]_{B_k}) \geq \alpha$.

Define 5. Let $S = (U, C \cup L, V)$ be a multi-label decision table. The coarse decision function $RDC^{v}_{(\alpha,\beta)}(x)$, the fine decision function $RDF^{v}_{(\alpha,\beta)}(x)$, and the uncertainty decision function $RDU^{v}_{(\alpha,\beta)}(x)$ are defined as follows:

$$RDC^{\upsilon}_{(\alpha,\beta)}(x) = \bigcup_{i=1}^{q} \{l_i \mid \rho_i^1 \geqslant \upsilon\}, l_i \in L, \tag{28}$$

$$RDF^{\upsilon}_{(\alpha,\beta)}(x) = \bigcup_{i=1}^{q} \{l_i \mid \rho_i^2 \geqslant \upsilon\}, l_i \in L, \tag{29}$$

$$RDU^{v}_{(\alpha,\beta)}(x) = RDC^{v}_{(\alpha,\beta)}(x) - RDF^{v}_{(\alpha,\beta)}(x), \tag{30}$$

$$\rho_i^1 = \frac{|\, r_i^1(x)\,|}{m}, \quad \rho_i^2 = \frac{|\, r_i^2(x)\,|}{m}, \tag{31}$$

where υ represents the variable degree, $0 \leq v \leq 1$. m is the total number of features. ρ_i^1 denotes the proportion of all granules satisfying $\Pr(E_i \mid [x]_{B_k}) > \beta$ to all features. ρ_i^2 represents the proportion of all granules satisfying $\Pr(E_i \mid [x]_{B_k}) \geq \alpha$ to all features.

For any $x \in U$, the RDC function represents the set of all labels that satisfy $\rho_i^1 > v$ and the RDF function represents the set of all labels that satisfy $\rho_i^2 > v$. The RDF function is the set of labels that contain a strong correlation with the object x. The RDC

Table 3
The calculation result of Example 2.

$\upsilon = 0.25$	$\upsilon = 0.5$
$RDC^{0.25}_{(\alpha,\beta)}(x_1) = \{l_1, l_2, l_3\}$	$RDC^{0.50}_{(\alpha,\beta)}(x_1) = \{l_1, l_2, l_3\}$
$RDC^{0.25}_{(\alpha,\beta)}(x_2) = \{l_1, l_2, l_3\}$	$RDC^{0.50}_{(\alpha,\beta)}(x_2) = \{l_1, l_2, l_3\}$
$RDF^{0.25}_{(\alpha,\beta)}(x_1) = \{l_1, l_2\}$	$RDF^{0.50}_{(\alpha,\beta)}(x_1) = \{l_1\}$
$RDF^{0.25}_{(\alpha,\beta)}(x_2) = \{l_1, l_2, l_3\}$	$RDF^{0.50}_{(\alpha,\beta)}(x_2) = \{l_2\}$
$\upsilon = 0.75$	$\upsilon = 1$
$RDC^{0.75}_{(\alpha,\beta)}(x_1) = \{l_1, l_3\}$	$RDC^{1}_{(\alpha,\beta)}(x_1) = \{l_1\}$
$RDC^{0.75}_{(\alpha,\beta)}(x_2) = \{l_1, l_2, l_3\}$	$RDC^{1}_{(\alpha,\beta)}(x_2) = \{l_1, l_2\}$
$RDF^{0.75}_{(\alpha,\beta)}(x_1) = \{l_1\}$	$RDF^{1}_{(\alpha,\beta)}(x_1) = \{l_1\}$
$RDF^{0.75}_{(\alpha,\beta)}(x_2) = \{l_2\}$	$RDF^{1}_{(\alpha,\beta)}(x_2) = \{\}$

function is the set of labels that may correlate with the object x. The set of labels that is generated by the RDU function represents the uncertainty that the object x has. The labels in this set cannot be determined whether they are necessarily related to the object x or not.

A higher value of υ indicates a more stringent condition for the decision and a greater degree of pessimism. As the value of υ increases, the labels generated by the RDF function gradually flow into the set generated by the RDU function, and the labels generated by the RDU function flow into the set of irrelevant labels. The model degenerates to an optimistic MGDRS for $\upsilon = 0$ and to a pessimistic MGDRS for $\upsilon = j/m$.

In this context, we consider each feature as a granularity space. It is a common approach in constructing granularity spaces, where each feature is treated as an independent space, and the different values of the feature represent different states or variations of that feature within the data. By modelling each feature as a granularity space, we can better analyze the relationships between features. Specifically, if two features are associated with the same label or if they belong to the same RDC or RDF function, then these two features are considered to be correlated.

Example 2. According to the multi-label decision table given in Table 1, we calculate the RDC function and the RDF function for object x_1 and x_2 under different degree υ, and the results are shown as Table 3, where $\alpha = 0.6, \beta = 0.4$.

From Table 3, it can be seen that both the RDC function and the RDF function show a decreasing tend as the value of υ increases and the retained labels correlate more strongly with the object x. As the value of υ increases, for the RDC function, the labels need to be correlated with more features in order to satisfy the conditions of the RDC function, whereas for the RDF function, the labels need to be strongly correlated with more features in order to satisfy the conditions of the RDF function. Therefore, Setting a reasonable value for the hyperparameter υ is the key to the success of the proposed algorithm. A high value of υ can lead to decision condition that is too pessimistic, causing some relevant features to be treated as irrelevant and eliminated, while a low value of υ can result in overly optimistic decision conditions, causing many redundant features to be selected without achieving the goal of dimensionality reduction.

This can also result in the retention of some redundant features if the value of υ is inappropriate. As the above calculation results shown, when υ is taken as 0.5 or 0.75, it is obvious that both the RDF function and RDC function contain values, and there are obvious boundaries between them, which portray the certainty and uncertainty more clearly. Therefore, the appropriate value of υ can better characterize each object.

Define 6. Let $S = (U, C \cup L, V)$ be a multi-label decision table, For an arbitrary $X \subseteq U$, we can define the label-upper approximation $\overline{RDPL}^{\upsilon}_{(\alpha,\beta)}(X)$, the label-down approximation $\underline{RDPL}^{\upsilon}_{(\alpha,\beta)}(X)$ and the label boundary $BAND^{\upsilon}_{(\alpha,\beta)}(X)$ of S as follows:

$$\overline{RDPL}^{\upsilon}_{(\alpha,\beta)}(X) = \bigcup_{x \in X} RDC^{\upsilon}(x), \tag{32}$$

$$\underline{RDPL}^{\upsilon}_{(x,\beta)}(X) = \bigcap_{x \in X} RDF^{\upsilon}(x), \tag{33}$$

$$BAND^{\upsilon}_{(\alpha,\beta)}(X) = \overline{RDPL}^{\upsilon}(X) - \underline{RDPL}^{\upsilon}(X). \tag{34}$$

The label-upper approximation is the set of labels that correlate with some of the objects, while the label-down approximation is the set of labels that are strongly correlated with all of the objects, and the label boundary is the set of labels that represents the uncertainty of the labels between the label-upper approximation and the label-down approximation.

Theorem 1. *Let $S = (U, C \cup L, V)$ be a multi-label decision table, for arbitrary $X \subseteq U$ and $B \subseteq C$, some theorems can be derived as follows:*

1. $\forall x \in U, RDF^{\upsilon}_{(\alpha,\beta),B}(x) \leq RDC^{\upsilon}_{(\alpha,\beta),B}(x),\ RDU^{\upsilon}_{(\alpha,\beta),B}(x) \leq RDC^{\upsilon}_{(\alpha,\beta),B}(x).$
2. $\forall x \in U, RDC^{\upsilon}_{(\alpha,\beta),B}(x) \leq RDC^{\upsilon}_{(\alpha,\beta),C}(x).$

3. $\forall x \in U, RDF^{\vee}_{(\alpha,\beta),B}(x) \leq RDF^{\vee}_{(\alpha,\beta),C}(x)$.
4. $\underline{RDPL}^{\vee}_{(\alpha,\beta),B}(X) \leq \overline{RDPL}^{\vee}_{(\alpha,\beta),B}(X)$.
5. $\underline{RDPL}^{\vee}_{(\alpha,\beta),B}(\phi) = \overline{RDPL}^{\vee}_{(\alpha,\beta),B}(\phi) = \phi$.

Proof 1.

1. From the fundamental theorem, it can be concluded that $RDF^{\vee}_{(\alpha,\beta),B}(x) \leq RDC^{\vee}_{(\alpha,\beta),B}(x), RDU^{\vee}_{(\alpha,\beta),B}(x) \leq RDC^{\vee}_{(\alpha,\beta),B}(x)$.
2. Since $B \subseteq C$, the attribute set C has more features to satisfy the RDC function than attribute set B. Thus, it is proved that $RDC^{\vee}_{(\alpha,\beta),B}(x) \leq RDC^{\vee}_{(\alpha,\beta),C}(x)$.
3. The reasoning is similar to that of item 2.
4. According to the previous derivation of $RDF^{\vee}_{(\alpha,\beta),B}(x) \leq RDC^{\vee}_{(\alpha,\beta),B}(x)$ and the definition of label-upper approximation and label-down approximation, it can be proved that $\underline{RDPL}^{\vee}_{(\alpha,\beta),B}(X) \leq \overline{RDPL}^{\vee}_{(\alpha,\beta),B}(X)$.
5. When $X = \phi$, it can be proved that $\underline{RDPL}^{\vee}_{(\alpha,\beta),B}(\phi) = \overline{RDPL}^{\vee}_{(\alpha,\beta),B}(\phi) = \phi$ according to the definition of label-upper approximation and label-down approximation.

3.4. VMFS: variable-degree MGDRS-based multi-label feature selection

In this subsection we discuss how to use variable-degree multi-granulation decision-theoretic rough sets for MLFS.

Define 7. Let $S = (U, C \cup L, V)$ be a multi-label decision table, For any feature subset B, $B \subseteq C$, the importance of the feature subset B to all objects is denoted by using label approximate accuracy as follows:

$$IMP_B = \frac{\sum_U card_B\left(RDF^{\vee}_{B,(x,\beta)}(x)\right) + \varphi(B)}{\sum_U card_B\left(RDC^{\vee}_{B,(x,\beta)}(x)\right) + \Phi(B)}, \tag{35}$$

$$\varphi(B) = \begin{cases} \lambda_1, & \forall x \in U, \sum_U card(\mathrm{RDF}^{\vee}_{B,(a,\theta)}(x)) = 0 \\ 0, & \exists x \in U, \sum_U card(\mathrm{RDF}^{\vee}_{B,(a,\theta)}(x)) > 0. \end{cases}, \tag{36}$$

$$\Phi(B) = \begin{cases} \lambda_2, & \forall x \in U, \sum_U card(\mathrm{RDC}^{\vee}_{B,(a,\theta)}(x)) = 0 \\ 0, & \exists x \in U, \sum_U card(\mathrm{RDC}^{\vee}_{B,(a,\theta)}(x)) > 0 \end{cases}, \tag{37}$$

where λ_1 and λ_2 are two constants and $\lambda_1 \leq \lambda_2$. In this paper, λ_1 is set to 0.1 and λ_2 is set to 0.11 respectively. $card(\bullet)$ is a cardinality function representing the number of sets. The label approximation accuracy reflects the dependency and uncertainty of the label set L on the feature subset B.

Define 8. Let $S = (U, C \cup L, V)$ be a multi-label decision table. For any feature $c \in C$, the dependency of that attribute denoted as follows:

$$R(c) = IMP_C - IMP_{C-\{c\}}. \tag{38}$$

If $R(c) > 0$, it means that the deletion of feature c would lead to the loss of strong relation information between objects and labels within the RDF function, so the feature c should be retained. If $R(c) < 0$, it means that removing of feature c would result in the loss of the relation information between objects and labels within the RDU function, the feature c should also be kept. Whether there is information loss in RDU or RDF functions, the remaining subset of features still cannot accurately induce the equivalence relationships within the multi-label data, resulting in the inability to select useful features. Consequently, features that satisfy $|R(c)| > 0$ are retained.

The algorithm of multi-label feature selection based on variable-degree MGDRS is denoted as Algorithm 1. This algorithm can be divided into two primary phases. In the first phase, it calculates the dependencies between different features. Then, in the subsequent phase, the features which fulfil the specific criteria are selected and merged into the selected feature set. The time complexity of the first phase is $O(|U| \bullet |C| \bullet |L|)$. For the second phase, it is $O(|C|)$. So, the overall time complexity of the algorithm is $O(|U| \bullet |C| \bullet |L| + |C|)$.

Algorithm 1: Multi-label feature selection based on variable-degree MGDRS(VMFS).

Input: $\alpha, \beta, v, S = (U, C \cup L, V)$. //$\alpha$ and β are two thresholds for three-way decisions. v is a variable degree. //S is a multi-label decision table
Output: *reduct*. //*reduct* is the result of feature selection
Step 1: //compute the dependency of each $c \in C$
compute $IMP_C = \frac{\sum_U card_C\left(RDF^v_{C,(x,\beta)}(x)\right)+\varphi(C)}{\sum_U card_C\left(RDC^v_{C,(x,\beta)}(x)\right)+\Phi(C)}$;
foreach c *in* C **do**
  compute $IMP_{C-\{c\}} = \frac{\sum_U card_{C-\{c\}}\left(RDF^v_{C-\{c\},(x,\beta)}(x)\right)+\varphi(C-\{c\})}{\sum_U card_{C-\{c\}}\left(RDC^v_{C-\{c\},(x,\beta)}(x)\right)+\Phi(C-\{c\})}$;
  compute $R(c) = IMP_C - IMP_{C-\{c\}}$;
Step 2: //Calculate the *reduct* obtained from feature selection
$reduct = \phi$;
foreach c *in* C **do**
  if $|R(c)| > 0$ **then**
    $reduct \bigcup c \rightarrow reduct$;
return *reduct*;

Table 4
Multi-label datasets.

Name	Instances	Attribute	Label	Train	Test
emotion	592	72	6	391	202
birds	645	260	19	322	323
yeast	2417	103	14	1499	918
CHD_49	555	68	6	333	222
CAL500	502	59	174	251	251
flag	194	19	7	129	65
image	2000	294	5	1000	1000

4. Experiments

4.1. Experimental preparation

In order to validate the effectiveness of the proposed algorithm, experiments were conducted on six multi-label datasets from diverse fields, all of which could be obtained from the Mulan Library, and their details are shown in Table 4.

The *emotion* dataset is a small collection of music that categorises the emotions expressed by music according to the Tellegen-Watson-Clark emotion model. The *bird* dataset aims to identify different bird species. The *yeast* dataset comprises microarray expression and phylogenetic profiles for 2417 yeast genes, and each gene is labelled based on its functional category. The CHD_49 dataset contains information on coronary heart disease from traditional Chinese medicine, with 49 features filtered and selected by experts. The $CAL500$ dataset involves 502 songs annotated by the annotators according to various semantic categories such as instrumentation, vocal characteristics, genres, emotions, acoustic quality, etc., with a total of 174 labels. The $flag$ dataset focuses on the flags of various countries and related information with the objective of predicting certain attributes. The *image* dataset contains 2000 images with 294 features extracted by experts and a total of five possible category labels.

Five evaluation metrics [34] for multi-label algorithms are employed for the experiments, namely Average Precision, Hamming Loss, Coverage, One-error and Ranking Loss. Let $T = \{(x_i, Y_i^*) \mid 1 < i < n\}$ is a testing set and $y_i = \{y_{i1}, y_{i2}, \ldots y_{iq}\}$ represents the label matrix of the predicted results for each object x_i. y_{ij} is the jth label of the ith instance. If $y_{ij} = 1$, it means that the category label j of object i is positive, while if $y_{ij} = 0$, it means that the category label j of object i is negative. All multi-label evaluation metrics appearing in this paper are represented as follows:

Average Precision (AP): This metric evaluates situations where the category labels appearing before the relevant labels in the sorted sequence of category labels for a particular are still considered relevant.

$$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|Y_i^*\right|}\sum_{y_i \in Y_i^*}\frac{\left|\left\{y_i' \mid R_j\left(x_i, y_i'\right) \le R_f\left(x_i, y_i\right), y_i' \in Y_i^*\right\}\right|}{R_f\left(x_i, y_i\right)}, \tag{39}$$

where $R_f(\bullet, \bullet)$ is a ranking function. The larger the metric is the better. The optimal value is $AP = 1$.

Hamming Loss (HL): This evaluation measure assesses the incorrect assignment of objects. In other words, relevant labels not appearing in the predicted set of labels or irrelevant labels appear in the predicted set of labels.

$$HL = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\left|classifier(x_i)\nabla Y_i^*\right|, \tag{40}$$

where $classifier(\bullet)$ represents a multi-label classifier. ∇ used to calculate the difference between two sets (symmetric difference). The smaller the metric is the better. The optimal value is $HL = 0$.

Coverage (CV): This evaluation metric is used to measure the depth of search required to cover all relevant labels in the sorted sequence of category labels of an instance.

$$CV = \frac{1}{n}\sum_{i=1}^{n} MAX_{y_i \in Y^*} R_f(x_i, y_i) - 1, \tag{41}$$

where $R_f(\bullet,\bullet)$ is a ranking function. The smaller the value of the metric is the better.

One-error (OE): This evaluation metric is used to examine the case where the label at the top of the sequence does not belong to the set of relevant labels in the sorted sequence of category labels of the instance.

$$OE = \frac{1}{n}\sum_{i=1}^{n} \mathrm{II}\left[(ARGMAX_{y_i \in L} f(x_i, y_i)) \notin Y_i^*\right], \tag{42}$$

where II[•] indicates a judgement. It is 1 if the condition in the symbol is satisfied and 0 if it is not. The smaller the value of the metric is the better, with an optimal value of $OE = 0$.

Ranking Loss (RL): This evaluation metric is utilized to assess the presence of label misplacement in the sorted sequence of category labels for an instance, where irrelevant labels precede relevant ones.

$$RL = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\left|Y_i^*\right|\left|\bar{Y}_i^*\right|} \left|\left\{(y_i', y_i'') \mid f(x_i, y_i') \geq f(x_i, y_i'')\right\}\right|, (y', y'') \in \bar{Y}_i \times Y_i^*, \tag{43}$$

where $\overline{Y}_i^*$ is the complement of the set Y_i^* in the label space. The system's performance improves as the metric value decreases, and the optimal value is $RL = 0$.

For all features with values not in the range [0,1], they are normalized to between [0,1] and then the continuous values are discretized into triples. The data used for the classification test are only normalized, but not discretized. The multi-label classifier used in all experiments is MLkNN, which has a smoothing parameter $s = 1$ and a neighbourhood granularity $k = 10$. The operating environment for the experiment is Windows 10 with a 4-core 3.5 GHz Intel E5-1600 v3 CPU and 16 GB RAM.

4.2. Experimental results and discussion

Before conducting the experiments, some key parameters involved in the proposed algorithm need to be clarified. Firstly, it is important to determine several three-way decision thresholds for the variable degree MGDRS. They can be calculated by Equation (3), but here they are set to constant values, $\alpha = 0.6, \beta = 0.4$. Then, the other parameter, namely the variable degree υ, needs to be determined. In order to specify the optimal value of the parameter υ, we discuss the impact of the variable degree υ on the performance of the proposed algorithm on two datasets, the *emotion* and the CHD_49, using the MLkNN as the basic multi-label classifier. The smothing covariance s of MLkNN algorithm is set to 1, and the neighbourhood granularity k is set to 10. Then, υ takes a value between 0 and 1, while five evaluation metrics as well as the selected feature number n are varied with υ. The results of the experiments are shown in Fig. 3 and Fig. 4.

The blue curve represents the results of the original features without feature selection, while the red curve represents the results of feature selection at different variable degree υ. By comparing the experimental results in Fig. 3, it can be determined that the optimal value of υ for the *emotion* dataset should be between 0.1 and 0.2, as this is when AP performs best, with relatively small values of HL, OE, RL and CV and a relatively small number of selected features. Here we decide on the value of υ for *emotion* dataset is 0.17. For the CHD_49 dataset, we compare the experimental results of Fig. 4 and conclude that the optimal value of υ is in the range of 0.12-0.2. The value of AP in this range is the highest. In this range, we want to keep the HL, RL, OE, and CV as low as possible and the number of selected features as low as possible. We finally decided on a suitable value of υ is 0.12. The remaining datasets *yeast*, *flags*, *cal*500, *birds*, and *image* can be used in the same way to derive the appropriate υ values, which are 0.02, 0.68, 0.35, 0.12, and 0.02, respectively.

After determining the values of the parameters, the optimal feature selection results could be obtained on each dataset with VMFS. We sort the feature selection results by the magnitude of the absolute value of the dependency and then add two features to the classifier for classification in order each time. And the change of each evaluation metric (AP, HL, CV, RL, OE) is plotted as a curve. Taking three datasets *emotion*, *bird* and *yeast* as examples, the change of each evaluation index is shown in Fig. 5.

By analysing the results of the feature selection experiments on three multi-label datasets, as shown in Fig. 5, it can be seen that five performance metrics of the multi-label classification algorithm improve to varying degrees as the number of selected features increases. Among them, the *emotion* dataset exhibits the most obvious change, with the average precision (AP) increasing from around 0.58 to around 0.80. Additionally, the Hamming Loss (HL), the Overall Error (OE), the Ranking Loss (RL), and the Coverage (CV) all decrease significantly, with HL decreasing from about 0.32 to about 0.23, OE decreasing from about 0.58 to about 0.23, RL decreasing from about 0.44 to about 0.17, and CV decreasing from about 3.2 to about 1.2. The performance metrics on the *bird* dataset varies less significantly, yet still achieves notable improvements, except for HL due to its initially small value. The *yeast* dataset exhibits the smallest changes on three datasets, with minimal and incremental effects on each metric as features are added.

In order to further validate the effectiveness of the VMFS algorithm, we compare it with some classical multi-label feature selection algorithms on seven datasets, including MDDM_proj [5], MLNB [14] MDDM_spc [5], PMU [35], RF-ML [36], and the experimental results are shown in Tables 5–11. The parameter μ of MDDM_proj and MDDM_spc is set to 0.5. MLNB selects 30% of the original

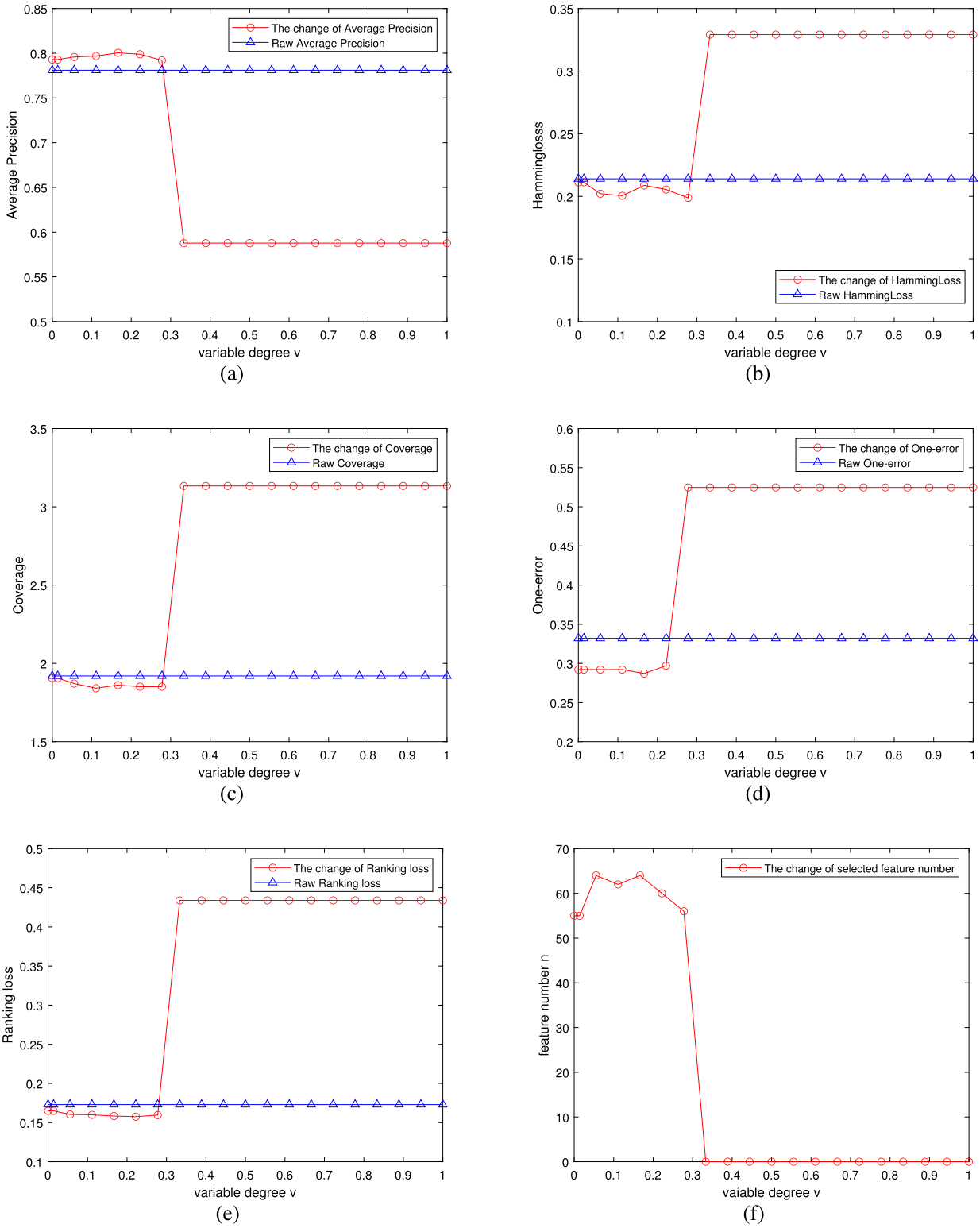


Fig. 3. The metrics change on the *emotion* dataset. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

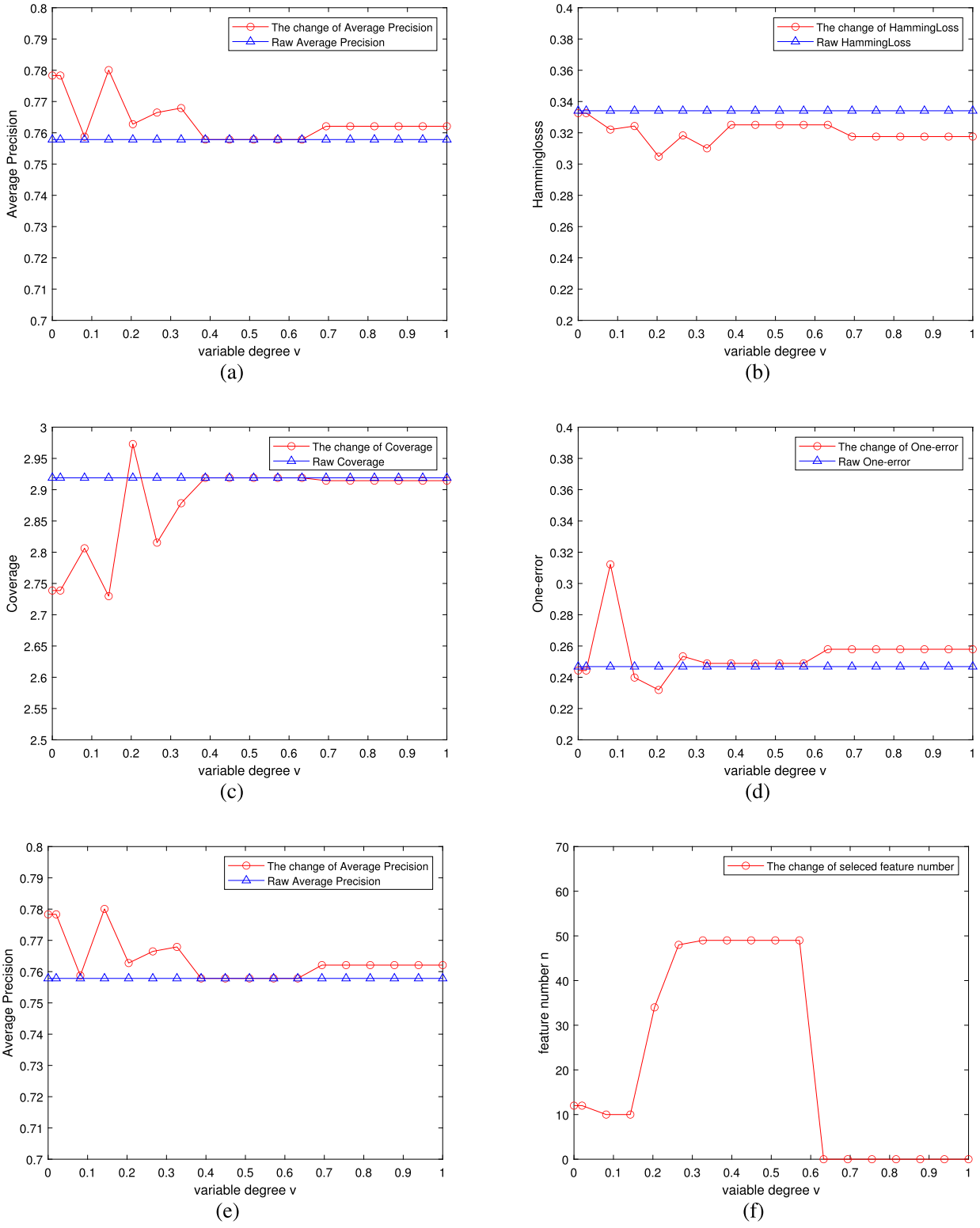
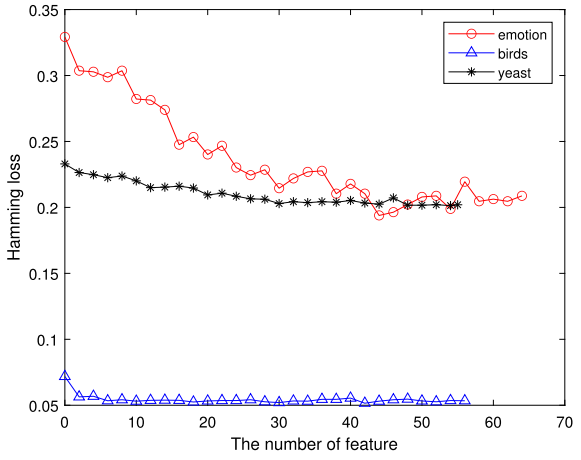
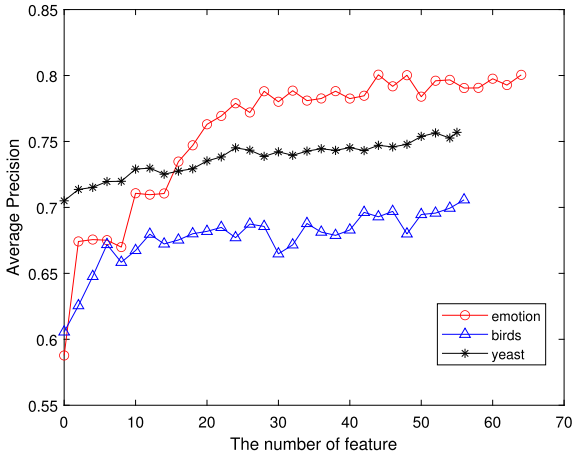


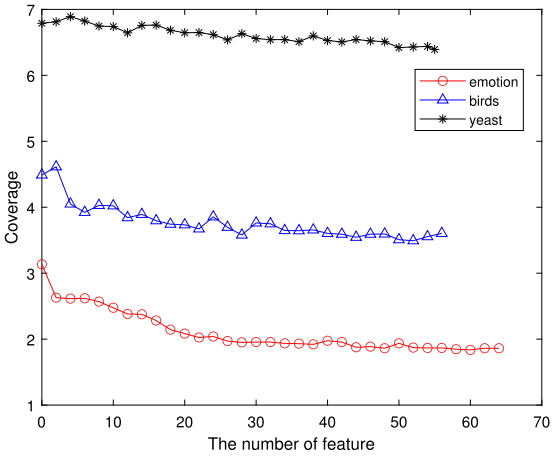
Fig. 4. The metrics change on the *CHD_49* dataset.



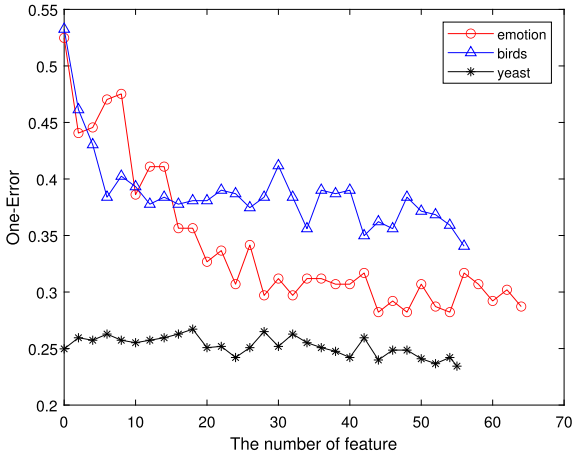
(a)



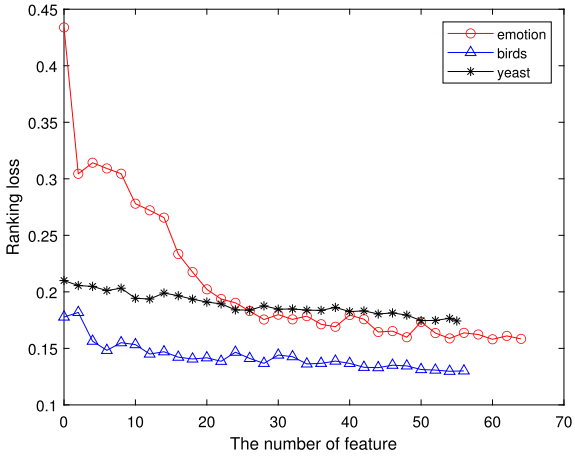
(b)



(c)



(d)



(e)

Fig. 5. The metrics change of three datasets with different selected feature number.

Table 5
Comparison of algorithm performance on *emotion* dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
64	VMFS	**0.800**	**1.842**	**0.209**	**0.158**	**0.287**
22	MLNB	0.784	1.975	0.245	0.179	0.302
3	MDDM_proj	0.782	1.970	0.278	0.185	0.317
6	MDDM_spc	0.761	2.020	0.247	0.195	0.357
64	PMU	0.788	1.906	0.204	0.174	0.312
43	RF-ML	0.736	2.213	0.267	0.230	0.381

Table 6
Comparison of algorithm performance on *yeast* dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
55	VMFS	**0.757**	**6.387**	**0.202**	**0.174**	**0.234**
31	MLNB	0.734	6.715	0.221	0.19	0.255
14	MDDM_proj	0.746	6.486	0.204	0.181	0.262
14	MDDM_spc	0.744	6.438	0.219	0.184	0.248
55	PMU	0.739	6.580	0.211	0.188	0.255
61	RF-ML	0.748	6.649	0.202	0.178	0.256

Table 7
Comparison of algorithm performance on $flag$ dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
13	VMFS	**0.825**	**3.708**	0.279	**0.205**	0.185
6	MLNB	0.809	3.769	0.290	0.231	0.211
3	MDDM_proj	0.800	3.831	0.321	0.223	0.246
4	MDDM_spc	0.815	3.723	0.301	0.214	**0.169**
13	PMU	0.797	3.815	0.325	0.250	0.215
11	RF-ML	0.800	3.846	**0.268**	0.221	0.246

Table 8
Comparison of algorithm performance on CHD_49 dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
9	VMFS	0.781	**2.761**	0.321	**0.218**	**0.240**
15	MLNB	0.777	2.806	**0.306**	0.229	0.244
2	MDDM_proj	**0.782**	2.842	0.318	0.226	0.249
5	MDDM_spc	0.765	2.815	0.339	0.247	0.272
9	PMU	0.760	2.910	0.307	0.245	0.262
29	RF-ML	0.767	2.905	0.318	0.238	0.258

Table 9
Comparison of algorithm performance on $CAL500$ dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
62	VMFS	0.492	**129.11**	**0.138**	0.184	**0.116**
21	MLNB	0.411	130.21	0.138	**0.180**	**0.116**
2	MDDM_proj	0.485	132.06	0.138	0.190	0.143
5	MDDM_spc	0.491	129.50	0.139	0.185	0.120
62	PMU	**0.493**	129.25	0.138	0.182	0.124
40	RF-ML	0.484	131.25	0.140	0.185	**0.116**

feature dimension as the number of retained features. The number of features returned by PMU is the same as our algorithm VMFS. RF-ML returns the top 60% features of the feature ranking results as the feature selection result. For each evaluation metric, a "↑" symbol indicates that a higher value is better, while a "↓" symbol indicates that a lower value is better. The optimal value of each evaluation metric is shown in bold form and the Num represents the number of remain features.

As shown in Tables 5–11, the proposed multi-label feature selection algorithm VMFS significantly outperforms other compared algorithms in general. On some datasets, such as the *emotion* and *yeast*, the proposed algorithm VMFS achieves optimal experimental results on all evaluation metrics compared to other algorithms. On the other datasets, VMFS does not achieve optimal performance on all evaluation metrics, but it still performs well on most of the metrics, with only some metrics falling slightly below the optimal algorithm. For example, on the $flag$ dataset, VMFS excels in AP, CV, and RL metrics, while its HL value (0.278) is slightly lower than the top RF-ML (0.268) and the OE value (0.185) is slightly lower than the top MDDM_spc (0.169). Similarly, on the $CAL500$,

Table 10
Comparison of algorithm performance on *birds* dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
56	VMFS	0.706	3.603	**0.053**	0.130	**0.341**
78	MLNB	0.649	3.796	0.078	0.145	0.440
13	MDDM_proj	0.707	**3.433**	0.056	**0.123**	0.359
19	MDDM_spc	**0.708**	3.598	0.062	0.128	0.350
56	PMU	0.649	3.468	0.061	0.128	0.477
156	RF-ML	0.677	3.811	0.056	0.141	0.409

Table 11
Comparison of algorithm performance on *image* dataset.

Num	Method	AP(↑)	CV(↓)	HL(↓)	RL(↓)	OE(↓)
163	VMFS	0.781	1.023	**0.178**	0.186	0.338
89	MLNB	0.738	1.186	0.208	0.223	0.401
25	MDDM_proj	**0.787**	**0.989**	**0.178**	**0.176**	**0.333**
30	MDDM_spc	0.773	1.069	0.180	0.195	0.344
163	PMU	0.778	1.280	0.187	0.185	0.339
176	RF-ML	0.751	1.146	0.199	0.317	0.381

Table 12
The values of evaluation metrics.

	AP	CV	HL	RL	OE
τ_{x^2}	10.816	15.163	6.795	6.367	9.698
τ_F	2.684	4.586	1.446	1.334	2.300

Table 13
p-value and correction p-value.

	AP	CV	HL	RL	OE
p-value	0.055	0.010	0.236	0.272	0.084
correction p-value	0.1375	0.005	0.272	0.272	0.140

*CHD*_500, and *birds* datasets, VMFS performs optimally or sub-optimally depending on the specific metrics, but there is no significant difference between the compared algorithms on some metrics. Overall, the VMFS algorithm generally outperforms the other comparative algorithms.

The proposed algorithm achieves relatively mediocre results on a few datasets, such as the *emotion* dataset. This is due to the fact that different datasets have different influences on the performance of the algorithm. However, an algorithm cannot achieve the best result in all datasets. However, although only one of the evaluation metrics HL achieves the best result, the rest of the evaluation metrics also get the second-best result, which is also a relatively good result.

When we propose a new algorithm and want to know whether the proposed algorithm performs better compared to existing algorithms, we usually utilize the model performance evaluation method. Friedman test [37] is a widely used statistical test method for comparing the overall effectiveness of k algorithms across n datasets. The calculation formula is as follows:

$$\tau_{x^2} = \frac{12n}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right), \tau_F = \frac{(n-1)\tau_{x^2}}{n(k-1)-\tau_{x^2}}, \tag{44}$$

where r_i represents the average ranking result of the ith algorithm, n represents the number of datasets, and k represents the number of algorithms. In this experiment the significance level α is set to 0.1 and τ_{χ^2} satisfies the $k-1$ and $(k-1)(n-1)$ degree of freedom distributions. The values of five evaluation indicators values are shown in Table 12. When $\alpha = 0.1$, $F(5,30) = 2.05$, it can be indicated from Table 12 that the proposed algorithm ranks slightly higher than other algorithms in two metrics HL and RL, and the performance gap is not significant. Therefore, in these two evaluation metrics, we cannot reject the null hypothesis that all algorithms are equal. For the evaluation metrics AP, CV, and OE, under the Friedman test [37], the null hypothesis of the same performance of six algorithms is rejected, which indicates that there is a difference in the performance. P-values are calculated from the results in Table 12, and used to assess the probability of observing the data or more extreme situations under the given hypothesis. To mitigate the risk of inflated false positive rates, a subsequent correction was applied using the Benjamini-Hochberg FDR correction [38], as shown in Table 13, with FDR = 0.15. Based on the results presented in Table 13, it is evident that the outcomes of the Friedman test are consistent.

In order to further analyze the differences between six algorithms, a post-hoc test such as the Bonferroni-Dunn [39] test is required. It is a statistical method which is designed to determine if there is a significant difference in the performance of the

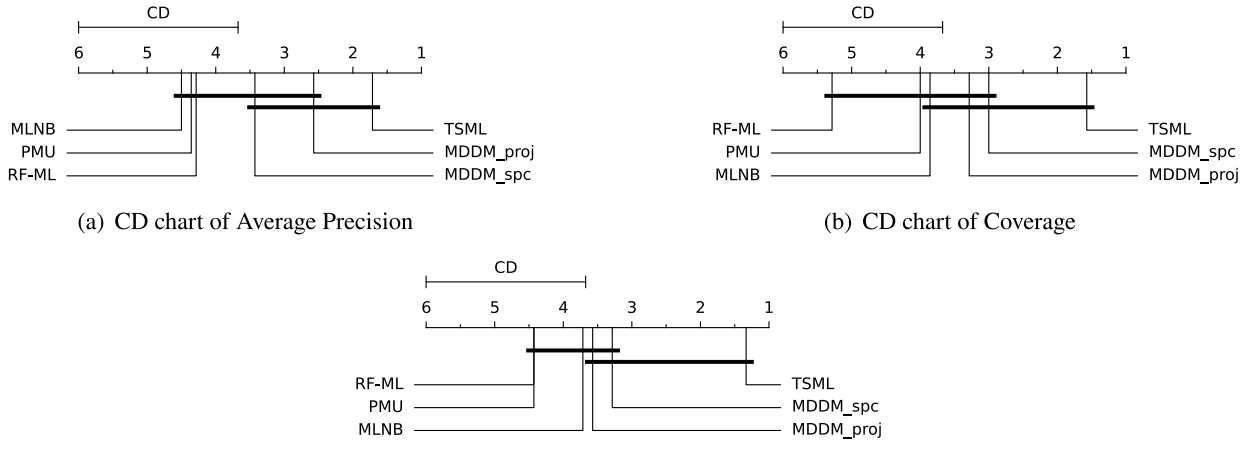


Fig. 6. CD bar of metrics.

evaluated algorithm compared to the remaining $k-1$ algorithms. It first calculates the difference in mean ranks between the algorithms, and then determines whether this difference is statistically significant. CD stands for the critical difference, which is expressed as follows:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}, \tag{45}$$

where q_α is the critical table value for the test and α is the significance level which is from reference [40]. For the Bonferroni-Dunn test, we have $q_\alpha = 2.326$ and $CD(k=6, n=7) = 2.326$. We visualize the results of the three parameters (AP, CV, RL) using a CD value chart, as shown in Fig. 6. The mean position of each algorithm in the ranking is displayed on the axes of the graph, where the ranking values increase progressively from left to right. When the average rank difference between the VMFS algorithm and the comparison algorithm is within one CD value, they are connected with a bold line, indicating that the comparison algorithm performs for different evaluation value metrics. If any algorithm is connected, it means that there is no significant difference between them, otherwise, it means that they are significantly different from each other.

From Fig. 6, it can be seen that for the evaluation metric AP, the performance gap between the three algorithms, VMFS, MDDM_proj and MDDM_spc, is within the value of one CD while the performance gap between the five algorithms, PMU, RF-ML, MLNB, MDDM_proj and MDDM_spc, is within the value of one CD. For the evaluation metric CV, the performance gap between the four algorithms, VMFS, MLNB, MDDM_proj, and MDDM_spc, is within one CD value while the performance gap between the five algorithms, PMU, RF-ML, MLNB, MDDM_proj, and MDDM_spc, is within one CD value. For the evaluation metric OE, the performance gap between each algorithm is similar to the metric AP. Overall, the performance of VMFS is relatively similar to that of MDDM_proj and MDDM_spc.

5. Conclusions

In order to flexibly deal with the uncertainty involved in the multi-labelled data, we introduce a variable degree for MGDRS and define coarse decision function, fine decision function and uncertainty decision function for multi-label learning. Then we define the upper and lower approximation of labels based on these decision functions and propose a MLFS algorithm based on variable degree MGDRS. Finally, we verified the effectiveness of the proposed algorithm through a series of comparison tests. However, we still encountered some problems in our experiments. First, the performance of feature selection in high-dimensional feature space is still poor, and the time to select features is too long. Second, the interdependence between labels is not fully considered in this paper. Therefore, we will continue to improve these aspects in our future work.

CRediT authorship contribution statement

Ying Yu: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – review & editing. **Ming Wan:** Validation, Writing – original draft. **Jin Qian:** Writing – review & editing. **Duoqian Miao:** Conceptualization, Supervision. **Zhiqiang Zhang:** Data curation. **Pengfei Zhao:** Data curation.

Declaration of competing interest

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank the Editors for their kindly help and the anonymous referees for their valuable comments and helpful suggestions. The work is partially supported by the National Natural Science Foundation of China (Serial No. 62163016, 62066014), the Natural Science Foundation of Jiangxi Provincial (Serial No. 20212ACB202001), the open project of State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure (Grant No. HJGZ2023203), the foreign expert project of Ministry of Science and Technology (No. G2023022005L), and the Jiangxi Double Thousand Plan (No. JSXQ2019102088).

References

[1] Weiwei Liu, Haobo Wang, Xiaobo Shen, Ivor W. Tsang, The emerging trends of multi-label learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 7955–7974.
[2] Gengyu Lyu, Songhe Feng, Yidong Li, Noisy label tolerance: a new perspective of partial multi-label learning, Inf. Sci. 543 (2021) 454–466.
[3] Wenbin Qian, Jintao Huang, Fankang Xu, Wenhao Shu, Weiping Ding, A survey on multi-label feature selection from perspectives of label fusion, Inf. Fusion 100 (2023) 101948.
[4] Wenbin Qian, Yanqiang Tu, Jin Qian, Wenhao Shu, Partial multi-label learning via three-way decision-based tri-training, Knowl.-Based Syst. 276 (2023) 110743.
[5] Yin Zhang, Zhi-Hua Zhou, Multilabel dimensionality reduction via dependence maximization, ACM Trans. Knowl. Discov. Data 4 (3) (2010) 1–21.
[6] Liang Sun, Shuiwang Ji, Jieping Ye, Multi-Label Dimensionality Reduction, CRC Press, 2013.
[7] Ping Zhang, Wanfu Gao, Juncheng Hu, Yonghao Li, Multi-label feature selection based on the division of label topics, Inf. Sci. 553 (2021) 129–153.
[8] Yuling Fan, Jinghua Liu, Wei Weng, Baihua Chen, Yannan Chen, Shunxiang Wu, Multi-label feature selection with local discriminant model and label correlations, Neurocomputing 442 (2021) 98–115.
[9] Weiping Ding, Chin-Teng Lin, Zehong Cao, Deep neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping pso with nearest-neighbor memeplexes, IEEE Trans. Cybern. 49 (7) (2018) 2744–2757.
[10] An-Da Li, Bing Xue, Mengjie Zhang, Improved binary particle swarm optimization for feature selection with new initialization and search space reduction strategies, Appl. Soft Comput. 106 (2021) 107302.
[11] Hongbin Dong, Jing Sun, Tao Li, Rui Ding, Xiaohang Sun, A multi-objective algorithm for multi-label filter feature selection problem, Appl. Intell. 50 (2020) 3748–3774.
[12] Lin Sun, Lanying Wang, Weiping Ding, Yuhua Qian, Jiucheng Xu, Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets, IEEE Trans. Fuzzy Syst. 29 (1) (2020) 19–33.
[13] Jia Zhang, Zhiming Luo, Candong Li, Changen Zhou, Shaozi Li, Manifold regularized discriminative feature selection for multi-label learning, Pattern Recognit. 95 (2019) 136–150.
[14] Min-Ling Zhang, José M. Peña, Victor Robles, Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19) (2009) 3218–3229.
[15] Yaojin Lin, Qinghua Hu, Jinghua Liu, Jie Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[16] Yuling Fan, Jinghua Liu, Jianeng Tang, Peizhong Liu, Yaojin Lin, Yongzhao Du, Learning correlation information for multi-label feature selection, Pattern Recognit. 145 (2024) 109899.
[17] Zdzislaw Pawlak, Rough sets and decision tables, in: Symposium on Computation Theory, Springer, 1984, pp. 187–196.
[18] Andrea Campagner, Davide Ciucci, Eyke Hüllermeier, Rough set-based feature selection for weakly labeled data, Int. J. Approx. Reason. 136 (2021) 150–167.
[19] Lin Sun, Shanshan Si, Weiping Ding, Xinya Wang, Jiucheng Xu. Tfsfb, Two-stage feature selection via fusing fuzzy multi-neighborhood rough set with binary whale optimization for imbalanced data, Inf. Fusion 95 (2023) 91–108.
[20] Yi Kou, Guoping Lin, Yuhua Qian, Shujiao Liao, A novel multi-label feature selection method with association rules and rough set, Inf. Sci. 624 (2023) 299–323.
[21] Jinghua Liu, Yaojin Lin, Weiping Ding, Hongbo Zhang, Cheng Wang, Jixiang Du, Multi-label feature selection based on label distribution and neighborhood rough set, Neurocomputing 524 (2023) 142–157.
[22] Wenbin Qian, Jintao Huang, Yinglong Wang, Yonghong Xie, Label distribution feature selection for multi-label classification with rough set, Int. J. Approx. Reason. 128 (2021) 32–55.
[23] Jie Duan, Q.H. Hu, L.J. Zhang, Y.H. Qian, D.Y. Li, Feature selection for multi-label classification based on neighborhood rough sets, J. Comput. Res. Dev. 52 (1) (2015) 56–65.
[24] Yaojin Lin, Qinghua Hu, Jinghua Liu, Jinkun Chen, Jie Duan, Multi-label feature selection based on neighborhood mutual information, Appl. Soft Comput. 38 (2016) 244–256.
[25] Hua Li, Deyu Li, Yanhui Zhai, Suge Wang, Jing Zhang, et al., A variable precision attribute reduction approach in multilabel decision tables, Sci. World J. (2014) 2014.
[26] Meishe Liang, Jusheng Mi, Tao Feng, Optimal granulation selection for multi-label data based on multi-granulation rough sets, Granul. Comput. 4 (2019) 323–335.
[27] Jiucheng Xu, Kaili Shen, Lin Sun, Multi-label feature selection based on fuzzy neighborhood rough sets, Complex Intell. Syst. 8 (3) (2022) 2105–2129.
[28] Yuhua Qian, Hu Zhang, Yanli Sang, Jiye Liang, Multigranulation decision-theoretic rough sets, Int. J. Approx. Reason. 55 (1) (2014) 225–237.
[29] Yiyu Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107–123.
[30] Eyke Hüllermeier, Sébastien Destercke, Ines Couso, Learning from imprecise data: adjustments of optimistic and pessimistic variants, in: Scalable Uncertainty Management: 13th International Conference, SUM 2019, Compiègne, France, December 16–18, 2019, Proceedings 13, Springer, 2019, pp. 266–279.
[31] Andrea Campagner, et al., Credal learning: weakly supervised learning from credal sets, Front. Artif. Intell. Appl. 372 (2023) 327–334.
[32] Yuanjian Zhang, Duoqian Miao, Witold Pedrycz, Tianna Zhao, Jianfeng Xu, Ying Yu, Granular structure-based incremental updating for multi-label classification, Knowl.-Based Syst. 189 (2020) 105066.
[33] Tianna Zhao, Yuanjian Zhang, Duoqian Miao, Intuitionistic fuzzy-based three-way label enhancement for multi-label classification, Mathematics 10 (11) (2022) 1847.
[34] Min-Ling Zhang, Zhi-Hua Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[35] Jaesung Lee, Dae-Won Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recognit. Lett. 34 (3) (2013) 349–357.
[36] Newton Spolaôr, Everton Alvares Cherman, Maria Carolina Monard, Huei Diana Lee, Relieff for multi-label feature selection, in: 2013 Brazilian Conference on Intelligent Systems, IEEE, 2013, pp. 6–11.

[37] Milton Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[38] Yoav Benjamini, Yosef Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc., Ser. B, Methodol. 57 (1) (1995) 289–300.
[39] Olive Jean Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (293) (1961) 52–64.
[40] Janez Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.