Knowledge-Based Systems 189 (2020) 105066

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Granular structure-based incremental updating for multi-label classification ☆

Yuanjian Zhang [a,b], Duoqian Miao [a,b,*], Witold Pedrycz [a,c,d], Tianna Zhao [a,b], Jianfeng Xu [a,b,e], Ying Yu [f]

[a] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, 201804, China
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6R 2V4
[d] System Research Institute, Polish Academy of Sciences, Warsaw, PL-01447, Poland
[e] Software College, Nanchang University, Jiangxi, 330047, China
[f] College of Software Engineering, East China Jiaotong University, Jiangxi, 330013, China

ARTICLE INFO

Article history:
Received 26 April 2019
Received in revised form 20 September 2019
Accepted 22 September 2019
Available online 16 October 2019

Keywords:
Incremental learning
Multi-label classification
Granular structure system
Three-way decisions

ABSTRACT

Incremental learning is an efficient computational paradigm of acquiring approximate knowledge of data in dynamic environment. Most of the research focuses on knowledge updating for single-label classification, whereas incremental mechanism for multi-label classification is of preliminary nature. This leads to considerable computation complexity to maintain desired performance. To address this challenge, we formulate a granular structure system (*GSS*). The proposed granular structure system in bottom-up way provides a systematic view on label-specific based classification. We demonstrate that the three-way selective ensemble (*TSEN*) model, a state-of-the-art solution for multi-label classification, is compatible with *GSS* in granulation. An incremental mechanism of *GSS* is introduced for both label-specific feature generation and optimization, and an incremental three-way selective ensemble algorithm for multiple instances immigration (*IMOTSEN*) is presented. Experiments completed on six datasets show that the proposed algorithm can maintain considerable classification performance while significantly accelerating the knowledge (*GSS*) updating.

1. Introduction

Granular Computing [1] is an effective structured problem solving methodology. It simulates human-centric operations realized in the presence of multifaceted data, and embraces a plethora of techniques which minimize uncertainty. The generic component, information granule, is rich in semantics and reflects a certain level of abstraction. As a representative model, rough set theory (RST), established by Pawlak [2] in 1982, is capable of dealing with ambiguous concept. Rough Sets have been extensively applied in various domains including sentiment analysis (e.g. [3]), social networks (e.g. [4]) and video analysis (e.g. [5]).

Previous studies usually assume a large collection of objects with all information is known in advance. However, the emerging features occur unexpectedly in a dynamic environment. To name a few examples, sensors continuously monitor the status of forest; navigation updates dynamically as the location changes; stock price fluctuates as new policy has been announced. A feasible solution is to assess the reliability of learnt knowledge incrementally beforehand, and make necessary modifications to accommodate information variations. For single-label learning, there are many inspiring works. Theoretically, knowledge updating can be induced by variations of objects, attributes, values, solely or simultaneously [6]. Leite et al. [7] investigated a fuzzy model-based control for nonlinear dynamic systems given object immigrations. Yu and Xu [8] proposed an incremental updating strategy for interval-valued ordered information system given different object variations. Luo et al. [9] discussed a matrix-based incremental mechanism with the consideration of decision risks when objects variations occur. Das et al. [10] enhanced the robustness of rough set-based pseudo outer-product model by incrementally updating the membership function. Xu et al. [11] extended object incremental mechanism to stream computing field. Yang et al. [12] explained the principles of attribute reduction updating with sample arriving based on fuzzy rough set. Lang et al. [13] presented a matrix-based approach for updating reduct of type-1 and type-2 characteristic matrices. Jing et al. [14] elaborated on the incremental mechanism of attribute reduction

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.105066.

* Corresponding author.
E-mail address: dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.knosys.2019.105066
0950-7051/

given object variations. Xie and Qin [15] presented an inconsistency degree driven mechanism for updating attribute reduction incrementally when objects and attributes vary simultaneously.

Multi-label learning [16,17] is an extension of single-label learning. Instances come with multiple labels present simultaneously, thus label correlations are introduced. Besides, an enormous number of features are collected to describe the semantics of label, which challenges the computational efficiency. The past decades witnessed a surge of developments on multi-label classification [18–23], and algorithms in static scenarios can be categorized into problem transformation (fit data to algorithm) and algorithm adaptation (fit algorithm to data). However, these approaches cannot be directly employed for cases of continuously generated pairwise instance-label. An alternative solution is to endow the model with incremental learning. Shi et al. [24] investigated an incremental updating strategy based on frequency of label combination. Class incremental operation is activated if the label combination is deemed as frequent, and instance incremental operation is activated otherwise. Lin et al. [25] considered the streaming label scenario, in which label information are available one by one, whereas the attribute side keeps unchanged. In contrast, Liu et al. [26] explored the streaming feature issue in multi-label classification, in which eligibility of group relevant features and inter-group features are subsequently examined for unchanged labels. Recently Zhu et al. [27] developed a learnware for emerging labels including three components named as optimizer, detector and classifier respectively. Nguyen et al. [28] proposed a Bayesian-based model by considering the pair-wise correlation within labels and relationship between features and labels simultaneously. In spite of the impressive work reported on incremental multi-label learning, the uncertainty of knowledge updating has not been critically investigated.

Rough set theory can describe the roughness of multi-label. Duan et al. [29] firstly analyzed the lower approximation preserved reduction of neighborhood rough set for multi-label classification. Xu et al. [30] suggested that integration of fuzzy rough set with learning label-specific features can generate a more desirable result. Lin et al. [31] focused on the integrity of label correlation for the usage of fuzzy rough set in multi-label classification. Li et al. [32] integrated fuzzy rough set with multiple kernel learning to enable the robustness of the model. The non-numerical case is discussed in [33], where a criterion called complementary decision reduction is defined. Despite these progress, two questions remain unsolved. The first problem is the concept definition. The co-existence of label combinations implies that at least one view is eligible for determining the association of a label with regard to a given instance. In other words, neither prediction as a whole nor concatenate label level prediction is plausible to be directly recognized as classification of included labels. The second problem is that the label association should be learned instead of stipulated. In a dynamic environment, the sparsity of labels implies that the determination of instance-label pair is challenging if a given concept representation is considered only. Thus, it is imperative to study the process of granulation in a systematical way.

Three-way Decisions (3WD) [34–37], also known as trisecting–acting–outcome (TAO) model, is a theoretical framework for reasoning with uncertainty. By introducing a third option, instances may be deferred for the affiliation of a particular knowledge structure. The law of three-way construction is versatile since it is independent of concrete algorithms. It is especially preferred when the multi-faceted information or pairwise granular structure is available [38–40]. For multi-label classification, the three-way selective ensemble model (TSEN) [41] is demonstrated to be superior than a collection of state-of-the-art methods. As new emerging labels are inevitable for limited known instances and difficult to update as a whole, we believe that the label-specific method should be extended to a dynamic environment. Our contributions are summarized as follows:

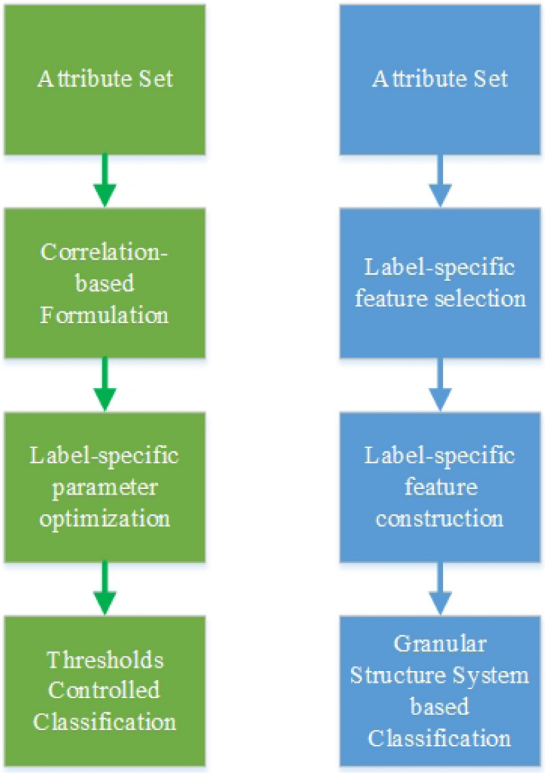


Fig. 1. Pipeline of multi-label learning with existing label-specific (left) and proposed algorithm (right).

- We firstly consider the problem of knowledge updating of multi-label under the framework of three-way decisions.
- The granulation of multi-label classification with label-specific view is formulated. The incremental mechanism at each granular layer is discussed.
- An incremental learning algorithm for multiple-objects immigration is developed. In most cases, the proposed algorithm can accelerate the construction of granular structure with comparable performance.

The comparison between the pipeline of existing label-specific multi-label learning and our algorithm is illustrated in Fig. 1. The incremental mechanism is implemented on the step "Label-specific feature selection" and "Label-specific feature construction". Comparing to the existing incremental model, the advantages of proposed model are twofold: (1) the order of label correlation is automatically determined by iteratively conducting pair-wise observations on feature representations, without the assumption of correlations among all labels; (2) the renewing of emerging label combinations is replaced by updating a group of label-specific knowledge, and it is possible to update the knowledge of labels selectively.

The paper is organized as follows. Section 2 outlines some preliminaries of the proposed method. Section 3 concerns the granulation of multi-label classification, and it provides the component for incremental extension. Section 4 elaborates on the details of incremental mechanism. Experiments and analyses of results are described in Section 5. Finally, we offer conclusions in Section 6.

2. Preliminaries

2.1. Three-way decisions

In this section, we review the basic notions and concepts for three-way decisions from the perspective of probability [35,42] defined in single-label learning paradigm.

Definition 1 ([43]). An information system is defined by a quadruple tuple: $IS = (U, A, V, f)$ where U is a finite non-empty set of data objects called universe; $A = C \bigcup D$ is a finite non-empty set of attributes, where C is a set of condition attributes, D is a set of decision attributes; $V = \bigcup\{V_a | a \in A\}$, where V_a is the set of values of attribute a; f is an information function from U to V, denoted as $f : U \times A \to V$.

Definition 2 ([43]). Given attributes set C, the $IND(C)$ denotes a binary relationship on IS with C, and is defined as:

$$IND(C) = \{(x, y) \in U \times U | \forall a \in C, f(x, a) = f(y, a)\}. \tag{1}$$

It can be easily shown that attributes C partition U into a number of non-overlapped sets. For simplicity, the sets of granules induced by $IND(C)$ is denoted as $[x]_C$. The conditional probability can be applied to measure the possibility of x belonging to concept X ($X \subset U$), and is defined as follows.

Definition 3 ([43]). Let $[x]_C$ be an equivalence class determined by x with respect to attribute C, then for an arbitrary subset $X \subset U$, we have the conditional probability as:

$$P(X | [x]_C) = \frac{|X \cap [x]_C|}{|[x]_C|}. \tag{2}$$

where $|\bullet|$ denotes the cardinality of a set.

For an ambiguous concept X, it can be approximated by a pair of operators as follows:

Definition 4 ([44]). Let $IS = (U, A, V, f)$ be an information system, the lower approximations of $X \subset U$ (denoted as $\underline{\mathcal{R}}_C^{\alpha}$) and upper approximation of $X \subset U$ (denoted as $\overline{\mathcal{R}}_C^{\beta}$) given attribute sets C are defined as follows:

$$\begin{aligned} &\underline{\mathcal{R}}_C^{\alpha} = \{x \in U \parallel P(X|[x]_C) \geq \alpha\}; \\ &\overline{\mathcal{R}}_C^{\beta} = \{x \in U | P(X|[x]_C) > \beta\}. \end{aligned} \tag{3}$$

From the perspective of decision-making, the result of conditional probability divides the whole universe into three regions named as positive region (POS), boundary region (BND) and negative region (NEG) respectively.

Definition 5 ([44]). Given a pair of thresholds α and β with $0 \leq \beta < \alpha \leq 1$, the positive region, boundary region and negative region are defined as follows:

$$\begin{aligned} &POS_C^{\alpha}(X) = \{x \in U \parallel P(X|[x]_C) \geq \alpha\}; \\ &BND_C^{(\alpha,\beta)}(X) = \{x \in U | \beta < P(X|[x]_C) < \alpha\}; \\ &NEG_C^{\beta}(X) = \{x \in U | P(X|[x]_C) \leq \beta\}. \end{aligned} \tag{4}$$

In order to balance the generality and reliability, a subset of attributes is selected. A wealth of uncertainty measures can be weighted to define a knowledge retention strategy, and one can resort to $\rho_B^{(\alpha,\beta)}(U/D)$ (where $B \subseteq C$, U/D denotes a partition of U induced by $IND(D)$), a general constraint-based form of reduct criterion over threshold pair (α, β). The selected attributes are jointly sufficient and individually necessary. Therefore, for $\rho_B^{(\alpha,\beta)}(U/D)$, (where $0 \leq \beta \leq \alpha \leq 1$) reduct criterion can be written as follows:

(1) $\rho_B^{(\alpha,\beta)}(U/D) = \rho_C^{(\alpha,\beta)}(U/D)$,

(2) $\rho_{B-\{a\}}^{(\alpha,\beta)}(U/D) \neq \rho_B^{(\alpha,\beta)}(U/D)$, for $\forall a \in B$.[1]

2.2. The TSEN model

TSEN [41] stands for three-way decisions with ensemble label correlation. It has been demonstrated to be effective in multi-label classification under static environment. In this section, we will review the main idea.

The model is based on two assumptions. Firstly, we assume that the characteristics with regard to an arbitrary label should be considered. Secondly, we assume that the knowledge representation of an arbitrary label can be improved by combining instance-based information with the relevant information granule. Three-way decisions are considered as the prediction framework to reduce uncertainty, whereas the selective ensemble is used to optimize the representation of each label. The notation $L = \{L_1, L_2, \ldots, L_m\}$, extends from D, is introduced to describe a label space with multiple labels. $L_i \in A$ holds since A stands for the non-empty set of attributes, i.e. $A = C \cup L$ holds for multi-label. To retain class-specific lower approximation of each label, for $b \in R_i$ and $R_i \subseteq C$, the corresponding reduct criterion for the ith label L_i, $\rho_B^{(\alpha,\beta)}(U/L_i)$, is equivalent to $\rho_{R_i}^{(\alpha_1,\alpha_2)}(U/L_i)$[2] and defined as:

$$\begin{aligned} &\underline{\mathcal{R}}_{R_i}^{\alpha_j}(Y_j) = \underline{\mathcal{R}}_C^{\alpha_j}(Y_j); \\ &\underline{\mathcal{R}}_{R_i-\{b\}}^{\alpha_j}(Y_j) \neq \underline{\mathcal{R}}_{R_i}^{\alpha_j}(Y_j). \end{aligned} \tag{5}$$

where $j = \{1, 2\}$, $Y_1 \subset U/L_i$ and $Y_2 \subset U/L_i$ represent the instances with/without label L_i. $0.5 \leq \alpha_1 \leq \alpha_2 \leq 1$. The significance of attribute $\gamma_{R_i}^{(\alpha_1,\alpha_2)}(U/L_i)$, is defined as:

$$\gamma_{R_i}^{(\alpha_1,\alpha_2)}(U/L_i) = 1 - (1 - \alpha_{R_i}^{(\alpha_1,\alpha_2)}(U/L_i)) \times GK(R_i). \tag{6}$$

where $\alpha_{R_i}^{(\alpha_1,\alpha_2)}(U/L_i) = \frac{\left|\underline{\mathcal{R}}_{R_i}^{\alpha_1}(Y_1)\right| + \left|\underline{\mathcal{R}}_{R_i}^{\alpha_2}(Y_2)\right|}{\left|\overline{\mathcal{R}}_{R_i}^{1-\alpha_1}(Y_1)\right| + \left|\overline{\mathcal{R}}_{R_i}^{1-\alpha_2}(Y_2)\right|}$ and $GK(R_i) = \frac{1}{|U|^2} \sum_{i=1}^{K} |X_i|^2$, $X_i \in U/R_i$.

The obtained information granule R_i constitutes the prototype of label-specific relevant features w.r.t. L_i. We assume that reductions with shared attributes, if they are computed by the same reduction strategy, signify stronger associations within labels. We introduce notation "$PWRL_i$" to specify the relevant label information of label L_i, denoted as:

$$PWRL_i = \bigcup \{L_j | R_i \bigcap R_j \neq \emptyset\}, \qquad \forall j \neq i. \tag{7}$$

Our next step is to develop an integration strategy to leverage the discernibility stemming from relevant feature representation. Instead of directly evaluating the similarity across those representations, we explore the relative tendency with regard to positive/negative classes. For each unseen instance x_j, we firstly define two repositories, P_i^j and N_i^j, to select the candidate positive/decision rules. Let R_i^k represent the kth relevant feature representation, $x_j^{R_i}$ represent the features of x_j on R_i, $R_i^k \to P$ and $R_i^k \to N$ represent instances with features of R_i^k is supposed to be positive class (with label L_i)/negative class (without label L_i) given the threshold pair (α_1, α_2), then P_i^j is defined as:

$$P_i^j = \bigcup_{R_i^k} \arg\max_{R_i^k} [\![R_i^k \to P]\!] \left[\!\left[Sim(x_j^{R_i}, R_i^k) > 0.5 \right]\!\right]. \tag{8}$$

[1] The second condition is a generalization of monotonicity $\rho_B^{(\alpha,\beta)}(U/D) \leq \rho_{B\cup\{b\}}^{(\alpha,\beta)}(U/D)$, for $\forall b \notin B$. The monotonicity is not always true.

[2] β can be omitted since it only works for determining upper approximation, herein we use (α_1, α_2) as superscript of ρ to indicate the criterion is controlled by (α_1, α_2). R_i substitutes B implies that reduct of L_i and L_j are different.

Analogously, we have N_i^j as:

$$N_i^j = \bigcup_{R_i^k} \arg\max_{R_i^k} [\![R_i^k \to N]\!] \left[\!\left[Sim(x_j^{R_i}, R_i^k) > 0.5 \right]\!\right]. \tag{9}$$

where $Sim(\cdot,\cdot)$ is Jaccard index,[3] $[\![\cdot]\!]$ is an indicator function and reaches 1 if condition $\cdot$ holds and reaches 0 otherwise.

The purpose for this selection operation is to find the corresponding features on R_i^k. In other words, an instance x_j is assumed to with label L_i if the representation on relevant labels (the features of x_j on the reduct of kth relevant label R_i^k, denoted as $x_j^{R_i^k}$) is included in a set (a feature set from the training set that comes from the kth relevant label of R_i and has the same representations as x_j on feature R_i, denoted as $C_{(i,j)}^k$) and is recognized as positive class in the sense of α_1 ($f(x_j^{R_i}) = 1$), and without label L_i if $x_j^{R_i^k}$ is included in $C_{(i,j)}^k$ and satisfies $f(x_j^{R_i}) = 0$. Regarding the importance of similarity with regard to all relevant features equally, we have positive label correlation degree lcp_i^j and negative label correlation degree lcn_i^j for x_j on label L_i.

$$lcp_i^j = \sum_k \frac{\left[\!\left[x_j^{R_i^k} \subseteq C_{(i,j)}^k \right]\!\right]}{|PWRL_i|}. \tag{10}$$

where $L_i^k \in PWRL_i \wedge f(x_j^{R_i}) = 1$.

$$lcn_i^j = \sum_k \frac{\left[\!\left[x_j^{R_i^k} \subseteq C_{(i,j)}^k \right]\!\right]}{|PWRL_i|}. \tag{11}$$

where $L_i^k \in PWRL_i \wedge f(x_j^{R_i}) = 0$.

Finally, we determine the label affiliation of L_i to x_j in the following way:

$$f_i(x_j) = \begin{cases} 1 & lcp_i^j > lcn_i^j, \\ 0 & lcp_i^j < lcn_i^j, \\ g_i(x_j) & otherwise. \end{cases} \tag{12}$$

where $g_i(x_j)$ is a function and determines the association of L_i to x_j via the features of R_i.

Since the reduction serves as the component for ensemble, our model is hierarchically developed. The overall complexity of *TSEN* is $O(|U/C \parallel C|^2|L|)$.

3. Granular structure system: an descriptor for multi-label

Definition 6 (*Multi-label Information System*)**.** A multi-label information system (MLIS) is defined by a quadruple tuple: $MLIS = (U, A, V, f)$ where U is a finite non-empty set of objects; $A = C \bigcup L$ is a finite non-empty set of attributes, where $C = \{c_1, c_2, \ldots, c_n\}$ is a set of condition attributes, $L = \{L_1, L_2, \ldots, L_m\}$ is a set of labels; $V = \bigcup\{V_a \,|a \in A\}$, where V_a is the set of values of attribute a; f is an information function from U to V, denoted as $f : U \times A \to V$.

To simulate the case of objects addition, we introduce the notation U^t to represent the instances at time t. For other elements in *MLIS* (i.e. A, V, f), we do not use the superscript t and assume that the underlying information, including the value domain, remains unchanged in the duration of updating. Typically, a brief example is provided:

[3] $Sim(x, y) = \frac{|x \cap y|}{|x \cup y|}$.

Table 1
An example of multi-label information system [33].

U^t	c_1	c_2	c_3	L_1	L_2	L_3
x_1	1	2	1	1	0	0
x_2	3	2	2	0	1	0
x_3	1	2	1	1	0	1
x_4	2	3	1	1	0	1
x_5	2	3	1	0	0	1
x_6	1	2	2	0	1	0
x_7	2	3	1	1	1	1
x_8	1	2	2	1	1	1
x_9	1	1	2	0	1	1
x_{10}	3	1	1	1	1	1
x_{11}	1	1	2	1	1	0

Example. A multi-label information system $MLIS^t = (U^t, A, V, f)$ at time t is given in Table 1, where $U^t = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}\}$, $A = C \cup L$, with $C = \{c_1, c_2, c_3\}$ and $L = \{L_1, L_2, L_3\}$. It can be observed that each object in U^t is associated with at least one label from L and that each label from L is associated with at least one object in U^t. It follows that $U^t/C = \{X_1, X_2, \ldots, X_6\}$, where

$X_1 = \{x_1, x_3\}, X_2 = \{x_2\}, X_3 = \{x_4, x_5, x_7\},$

$X_4 = \{x_6, x_8\}, X_5 = \{x_9, x_{11}\}, X_6 = \{x_{10}\}.$ □

Problem-transformation based algorithms constitute a large proportion of the multi-label learning algorithm. By completing problem transformation, original problems are converted potentially to a great number of well-established learning scenarios. Thus, the essence of the approach is twofold: (1) how to define problem transformation operations, and (2) how to synthesize the instance-based classification result.

Multi-class classification is one of the most favored problem transformation solutions in multi-label classification. Let $LS = \{LS_1, LS_2, \ldots, LS_i, \ldots, LS_r\}$ be a description of the label set L. For each LS_i, a multi-class subproblem is formulated. Assume an assemble of binary relations $Rel = \{Rel_1, Rel_2, \ldots, Rel_i \cdots, Rel_r\}$ is considered for corresponding label learning problems, we have information granule structure $GS = \{GS_1, GS_2, \ldots, GS_i, \cdots, GS_r\}$, where GS_i is determined by $MLIS, Rel_i, LS_i$ simultaneously, and $1 \le i \le r$. For different LS_i, the binary relations can be homogeneous or heterogeneous, depending on the nature of the classification problem. The granularity of GS_i may be different from GS_k, depending on the distribution of relevant attributes and labels. By solving each multi-class problem, the features of multi-label on a specific label L_i can be bottom-up constructed. For a dynamic environment, we arrive at the definition of granular correlation matrix.

Definition 7 (*Granular Correlation Matrix*)**.** $GM^t = (gm_{i,k}^t)_{j\times j}$ denotes a granular correlation matrix at time t, where $gm_{i,k}^t$ is the correlation metric regarding granular structure GS_i and GS_k at time t. The basic element $gm_{i,k}^t$ is defined as

$$gm_{i,k}^t = \frac{\left|\{b \mid b \in GS_i^t \cap GS_k^t\}\right|}{\left|\{b \mid b \in GS_i^t \cup GS_k^t\}\right|}. \tag{13}$$

Based on Definition 7, we have the definitions of granular structure system.

Definition 8 (*Granular Structure System*)**.** A granular structure system (GSS) at time t is defined by a quadruples tuple $GSS^t = (MLIS^t, Rel, GS^t, GM^t)$, where $MLIS^t = (U^t, A, V, f)$ is a multi-label information system, GS^t stands for granular structure determined by *Rel* at time t.

In what follows, the description of label set L is implemented in a label-specific way, denoted as $LS = \{LS_1, LS_2, \cdots, LS_m\}$, with $LS_i = L_i$. The basic element $gm^t_{i,j}$ in GM^t is thus defined as:

$$gm^t_{i,j} = \frac{|R^t_i \cap R^t_j|}{|R^t_i \cup R^t_j|}. \tag{14}$$

where R^t_i and R^t_j are the reduct of LS_i and LS_j at time t respectively.

It is easy to verify that cardinality of $gm^t_{i,j}$ is at least zero and at most one. This is implied by the fact that $\emptyset \subseteq R^t_i \cap R^t_j \subseteq R^t_i \cup R^t_j$, which implies $0 \leq gm^t_{i,j} = \frac{|R^t_i \cap R^t_j|}{|R^t_i \cup R^t_j|} \leq 1$.

Let $\theta^t_j(GS_j)$ be the indicator of instances on label(s) GS_j given feature representations on GS_j at time t. From the perspective of label, the information function f_L with the semantics $U \times L \rightarrow V_L$ can be rewritten as $f^t_L = \{f^t_{L_1}, f^t_{L_2}, \ldots, f^t_{L_m}\}$. Given an representation of GS, the $f^t_{L_i}$ can be determined by an aggregation function $agg(\cdot)$, i.e.,

$$f^t_{L_i} = agg(\theta^t_j(GS_j)). \tag{15}$$

where L_i is associated with GS_j. The aggregation function $agg(\theta^t_j(GS_j))$ can be implemented with majority, mean, or weighted sum etc for all included $\theta^t_j(GS_j)$. As to algorithms adopt binary relevance strategy, the $agg(\cdot) = \theta^t_j(\cdot)$ holds $\forall L_i$.

For $agg(L_i)$, we stipulate as the mean of relative label preferences w.r.t. positive/negative classes on all relevant granular structure $\theta^t(GS_j)$. i.e.,

$$f^t_{L_i} = \left[\left[\sum(\theta^t_j(GS_j) = 1) - \sum(\theta^t_j(GS_j) = 0) > 0\right]\right]. \tag{16}$$

where $L_j \in PWRL_i$.

4. Knowledge-based incremental updating for multi-label classification

Incremental knowledge updating is capable of reducing repetitive computations and renewing underlying structure when it is necessary. For convenience of description, we will firstly present the incremental mechanism and then summarize it as an incremental algorithm.

4.1. Incremental updating mechanism

We consider the incremental mechanism of GSS. Being specific, the mechanism concerns the updating of granular structure GS and granular correlation matrix GM. Given $0.5 \leq \alpha_1 \leq \alpha_2 \leq 1$,[4] the updating mechanism of GS and GM for multiple instances is elaborated in Sections 4.1.1 and 4.1.2 respectively.

4.1.1. Updating mechanism of granular structure

In our previous study [41], the lower approximation is selected as a reduct criterion. This criterion is capable of maintaining the approximation quality from available data, however it is an absolute criterion. In order to make a fair comparison with incremental algorithms, the criterion is replaced by a relative measure:

$$\rho^{(\alpha_1,\alpha_2)}_{R^t_i}(U^t/L_i) = \frac{|\underline{\mathcal{R}}^{\alpha_1}_{R^t_i}(Y^t_1)| + |\underline{\mathcal{R}}^{\alpha_2}_{R^t_i}(Y^t_2)|}{|U^t|}. \tag{17}$$

where Y^t_1 and Y^t_2 are the instances associated with label L_i and without label L_i at time t. The semantics of the renewed criterion is the preservation of lower approximation rate.

[4] We penalize more on misclassification of positive class since positive class tends to be misclassified more easily.

Example. Consider the multi-label information system given in Table 1 collected at time t. Given $(\alpha_1, \alpha_2) = (1, 1)$, we have:

$$\rho^{(\alpha_1,\alpha_2)}_C(U^t/L_1) = \frac{2+1+0+0+0+1}{11} = \frac{4}{11},$$
$$\rho^{(\alpha_1,\alpha_2)}_C(U^t/L_2) = \frac{2+1+0+2+2+1}{11} = \frac{8}{11},$$
$$\rho^{(\alpha_1,\alpha_2)}_C(U^t/L_3) = \frac{0+1+3+0+0+1}{11} = \frac{5}{11}.$$

Additionally, we compute the equivalence classes induced by U^t/c_1, U^t/c_2 and U^t/c_3 as:

$U^t/c_1 = \{\{x_1, x_3, x_6, x_8, x_9, x_{11}\}, \{x_4, x_5, x_7\}, \{x_2, x_{10}\}\}$,
$U^t/c_2 = \{\{x_9, x_{10}, x_{11}\}, \{x_1, x_2, x_3, x_6, x_8\}, \{x_4, x_5, x_7\}\}$,
$U^t/c_3 = \{\{x_1, x_3, x_4, x_5, x_7, x_{10}\}, \{x_2, x_6, x_8, x_9, x_{11}\}\}$.

For L_1, we examine the candidate reduction by taking addition–deletion strategy as:

$$\underline{\mathcal{R}}^{\alpha_1}_{c_1}(Y^t_1) = \emptyset, \underline{\mathcal{R}}^{\alpha_2}_{c_1}(Y^t_2) = \emptyset$$
$$\overline{\mathcal{R}}^{1-\alpha_1}_{c_1}(Y^t_1) = U^t, \overline{\mathcal{R}}^{1-\alpha_2}_{c_1}(Y^t_2) = U^t.$$
$$GK(c_1) = \frac{6^2+3^2+2^2}{11^2} = \frac{49}{121},$$
$$\gamma^{(\alpha_1,\alpha_2)}_{c_1}(U^t/L_1) = 1-(1-0)\times\frac{49}{121} = \frac{72}{121}.$$
$$\underline{\mathcal{R}}^{\alpha_1}_{c_2}(Y^t_1) = \emptyset, \underline{\mathcal{R}}^{\alpha_2}_{c_2}(Y^t_2) = \emptyset.$$
$$\overline{\mathcal{R}}^{1-\alpha_1}_{c_2}(Y^t_1) = U^t, \overline{\mathcal{R}}^{1-\alpha_2}_{c_2}(Y^t_2) = U^t.$$
$$GK(c_2) = \frac{3^2+5^2+3^2}{11^2} = \frac{43}{121},$$
$$\gamma^{(\alpha_1,\alpha_2)}_{c_2}(U^t/L_1) = 1-(1-0)\times\frac{43}{121} = \frac{78}{121}.$$
$$\underline{\mathcal{R}}^{\alpha_1}_{c_2}(Y^t_1) = \emptyset, \underline{\mathcal{R}}^{\alpha_2}_{c_2}(Y^t_2) = \emptyset.$$
$$\overline{\mathcal{R}}^{1-\alpha_1}_{c_2}(Y^t_1) = U^t, \overline{\mathcal{R}}^{1-\alpha_2}_{c_2}(Y^t_2) = U^t.$$
$$GK(c_3) = \frac{6^2+5^2}{11^2} = \frac{61}{121},$$
$$\gamma^{(\alpha_1,\alpha_2)}_{c_3}(U^t/L_1) = 1-(1-0)\times\frac{61}{121} = \frac{60}{121}.$$

We select c_2 since $c_2 = \arg\max_{c_i \in C} \gamma^{(\alpha_1,\alpha_2)}_{c_i}(U^t/L_1)$, where $i \in \{1, 2, 3\}$.

We need to add more attributes since $\rho^{(\alpha_1,\alpha_2)}_{c_2}(U^t/L_1) = 0 < \rho^{(\alpha_1,\alpha_2)}_C(U^t/L_1) = \frac{4}{11}$,

Then we consider the combination of $\{c_1, c_2\}$ and $\{c_2, c_3\}$.

Combination of $\{c_1, c_2\}$:

$$\underline{\mathcal{R}}^{\alpha_1}_{\{c_1,c_2\}}(Y^t_1) = \{x_{10}\}, \underline{\mathcal{R}}^{\alpha_2}_{\{c_1,c_2\}}(Y^t_2) = \{x_2\}.$$
$$\overline{\mathcal{R}}^{1-\alpha_1}_{\{c_1,c_2\}}(Y^t_1) = \{x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}\},$$
$$\overline{\mathcal{R}}^{1-\alpha_2}_{\{c_1,c_2\}}(Y^t_2) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{11}\}.$$
$$GK(\{c_1, c_2\}) = \frac{2^2+4^2+3^2+1^2+1^2}{11^2} = \frac{31}{121},$$
$$\gamma^{(\alpha_1,\alpha_2)}_{\{c_1,c_2\}}(U^t/L_1) = 1-(1-\frac{2}{20})\times\frac{31}{121} = \frac{1179}{1210}.$$

Combination of $\{c_2, c_3\}$:

$$\underline{\mathcal{R}}^{\alpha_1}_{\{c_2,c_3\}}(Y^t_1) = \{x_1, x_3, x_{10}\}, \underline{\mathcal{R}}^{\alpha_2}_{\{c_2,c_3\}}(Y^t_2) = \emptyset.$$
$$\overline{\mathcal{R}}^{1-\alpha_1}_{\{c_2,c_3\}}(Y^t_1) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}\},$$
$$\overline{\mathcal{R}}^{1-\alpha_2}_{\{c_2,c_3\}}(Y^t_2) = \{x_2, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}\}.$$
$$GK(\{c_2, c_3\}) = \frac{1^2+2^2+2^2+3^2+3^2}{11^2} = \frac{27}{121},$$
$$\gamma^{(\alpha_1,\alpha_2)}_{\{c_2,c_3\}}(U^t/L_1) = 1-(1-\frac{3}{19})\times\frac{27}{121} = \frac{2218}{2299}.$$

We select c_1 since $c_1 = \arg\max_{c_i \in C} \gamma_{c_i}^{(\alpha_1,\alpha_2)}(U^t/L_1)$, where $i \in \{1, 3\}$.

$\rho_{\{c_1,c_3\}}^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{3}{11} < \rho_C^{(\alpha_1,\alpha_2)}(U^t/L_1)$, Since $\rho_C^{(\alpha_1,\alpha_2)}(U^t/L_1)$ is $\frac{4}{11}$, we consider deletion of c_1, c_2, and c_3 in order of the value of $\gamma_{c_i}^{(\alpha_1,\alpha_2)}(U^t/L_1)$ from small to large.

Deletion of c_3:

$$\rho_{\{c_1,c_2\}}^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{5}{11} \neq \rho_C^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{4}{11}.$$

Deletion of c_1:

$$\rho_{\{c_2,c_3\}}^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{3}{11} \neq \rho_C^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{4}{11}.$$

Deletion of c_2:

$$\rho_{\{c_1,c_3\}}^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{4}{11} = \rho_C^{(\alpha_1,\alpha_2)}(U^t/L_1).$$

Therefore, we obtain $R_1^t = \{c_1, c_3\}$.

Similarly, for L_2, we have $R_2^t = \{c_2, c_3\}$, for L_3, we have $R_3^t = \{c_1, c_2\}$.

Given the unchanged attribute set, it will be time-consuming if we consider Definition 5 for the approximation set. An alternative solution is to revise the approximation set incrementally based on the variation of conditional probability, which is compared with a threshold pair (α_1, α_2). Let $Y_j^t \in U^t/L_k$ and $B \subseteq C$, $B_i^t \in U^t/C$ be the jth equivalence class of label L_k at time t, then given ith equivalence class w.r.t. attribute B, the variation trend for probability of Y_j^{t+1} conditioned on B_i^{t+1} at time $t+1$ (denoted as $P(Y_j^{t+1}|B_i^{t+1})$), as compared to probability of Y_j^t conditioned on B_i^t at time t (denoted as $P(Y_j^t|B_i^t)$), can be incrementally determined (see Lemma 1).

Lemma 1. *Let $Y_j^t \in U^t/L_k$ and $B_i^t \in U^t/B$ be the jth equivalence class of label L_k and ith equivalence class of attributes B at time t, ΔU^t are the immigrated instances at time $t+1$. $\Delta E_{Y_j}^{t+1} \in \Delta U^{t+1}/L_k$ and $\Delta E_{B_i}^{t+1} \in \Delta U^{t+1}/B$ be the immigrated instances with jth label in L_k and instances with ith feature vector at time $t+1$. Given $B_i^{t+1} \in U^t \cup \Delta U^{t+1}/B$ the variation trend for conditional probability $P(Y_j^{t+1}|B_i^{t+1})$ can be estimated as:*

$$\begin{aligned}
&P(Y_j^{t+1}|B_i^{t+1}) \geq P(Y_j^t|B_i^t), \quad \text{if } P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) \geq P(Y_j^t|B_i^t);\\
&P(Y_j^{t+1}|B_i^{t+1}) < P(Y_j^t|B_i^t), \quad \text{if } P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) < P(Y_j^t|B_i^t);\\
&P(Y_j^{t+1}|B_i^{t+1}) = P(Y_j^t|B_i^t), \quad \text{if } \Delta E_{B_i}^{t+1} = \emptyset.
\end{aligned} \tag{18}$$

Lemma 2. *Let $Y_j^t \in U^t/L_k$, $B_i^t \in U^t/B$ be the ith equivalence class of feature set at time t, $\Delta E_{B_i}^{t+1} \in \Delta U^{t+1}/B$ is the incremental instance assemble with feature vector $\Delta E_{B_i}^{t+1}$ at time $t+1$. Given thresholds α_k ($0.5 \leq \alpha_j \leq 1$), the lower approximation with regard to Y_j at time $t+1$ ($\underline{\mathcal{R}_B}(Y_j^{t+1})$) can be estimated as:*

$$\begin{aligned}
&\underline{\mathcal{R}_B}(Y_j^{t+1}) = \underline{\mathcal{R}_B}(Y_j^t) \cup \Delta E_{B_i}^{t+1} \quad \text{if } P(Y_j^{t+1}|B_i^{t+1}) \geq \alpha_j \\
&\qquad\qquad \wedge P(Y_j^t|B_i^t) \geq \alpha_j;\\
&\underline{\mathcal{R}_B}(Y_j^{t+1}) = \underline{\mathcal{R}_B}(Y_j^t) - B_i^t \quad \text{if } P(Y_j^{t+1}|B_i^{t+1}) < \alpha_j \\
&\qquad\qquad \wedge P(Y_j^t|B_i^t) \geq \alpha_j;\\
&\underline{\mathcal{R}_B}(Y_j^{t+1}) = \underline{\mathcal{R}_B}(Y_j^t) \cup B_i^t \cup \Delta E_{B_i}^{t+1} \quad \text{if } P(Y_j^{t+1}|B_i^{t+1}) \geq \alpha_j \\
&\qquad\qquad \wedge P(Y_j^t|B_i^t) < \alpha_j;\\
&\underline{\mathcal{R}_B}(Y_j^{t+1}) = \underline{\mathcal{R}_B}(Y_j^t) \quad \text{if } \Delta E_{B_i}^{t+1} = \emptyset \vee \\
&\qquad\qquad (P(Y_j^{t+1}|B_i^{t+1}) < \alpha_j \wedge P(Y_j^t|B_i^t) < \alpha_j).
\end{aligned} \tag{19}$$

Detailed proofs of Lemmas 1 and 2 can be found in appendix.

Table 2
Immigrated objects of multi-label information system at time $t+1$.

ΔU^{t+1}	c_1	c_2	c_3	L_1	L_2	L_3
x_{12}	1	1	2	1	1	0
x_{13}	3	2	1	0	1	1
x_{14}	1	1	1	1	0	1
x_{15}	2	2	1	1	0	1

Remark 1. The condition $0.5 \leq \alpha_j \leq 1$ implies that an instance x will not belong to the lower approximation of Y_k at time t, i.e. $\underline{\mathcal{R}_B}(Y_k^t)$, if it belongs to the lower approximation of Y_j at time t, i.e. $\underline{\mathcal{R}_B}(Y_j^t)$, and vice versa.

When a collection of objects (ΔU^{t+1}) is available to the decision system at time $t+1$, the next objective is to determine whether the approximation quality w.r.t. label at time t ($\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i)$) is preserved at time $t+1$ ($\rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$). To accelerate the calculation of $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$, we present the corresponding incremental mechanism.

Theorem 1. *Let $MLIS^t = (U^t, A = C \bigcup L, V, f)$ be a multi-label information system at time t, and ΔU^{t+1} denotes a new assemble of instances at time $t+1$. Suppose R_i^t is the reduct of L_i at time t, and $\Delta U^{t+1} = \Delta U_1^{t+1} \cup \Delta U_2^{t+1}$, $U^{t+1} = U^t \cup \Delta U^{t+1}$, $U^{t+1}/R_i^t = \{E_1^t, E_2^t, \ldots, E_q^t, \Delta E_{q+1}^{t+1}, \ldots, \Delta E_{q+s}^{t+1}, \Delta E_{q+s+1}^{t+1}, \ldots, \Delta E_{q+s+v}^{t+1}\}$, where $\Delta U_1^{t+1}/R_i^t = \{\Delta E_{q+1}^{t+1}, \Delta E_{q+2}^{t+1}, \ldots, \Delta E_{q+s}^{t+1}\}$, $\Delta U_2^{t+1}/R_i^t = \{\Delta E_{q+s+1}^{t+1}, \Delta E_{q+s+2}^{t+1}, \ldots, \Delta E_{q+s+v}^{t+1}\}$. For the equivalence classes in U^{t+1}/R_i^t, the first q equivalence classes represent the information granules which remain unchanged, the middle s equivalence classes represent the equivalence classes that can be merged with existing classes in U^t/R_i^t, and the remaining v equivalence classes represent the equivalence classes that cannot be combined with the existing equivalence classes deduced by U^t/R_i^t. Given two thresholds α_1 and α_2 with $0.5 \leq \alpha_1 \leq \alpha_2 \leq 1$, then approximate quality $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i)$ can be updated as:*

$$\begin{aligned}
&\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i)\\
&= \frac{\sum_{r=1}^{q} \left|\bigcup\{E_r^t \mid E_r^t \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})\}\right|}{|U^t| + \left|\Delta U^{t+1}\right|}\\
&\quad + \frac{\sum_{r=q+1}^{q+s} \left|\bigcup\{E_r^t \cup \Delta E_r^{t+1} \mid E_r^t \cup \Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})\}\right|}{|U^t| + \left|\Delta U^{t+1}\right|}\\
&\quad + \frac{\sum_{r=q+s+1}^{q+s+v} \left|\bigcup\{\Delta E_r^{t+1} \mid \Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})\}\right|}{|U^t| + \left|\Delta U^{t+1}\right|}.
\end{aligned} \tag{20}$$

where $j \in \{1, 2\}$.

The proof of Theorem 1 can be found in the appendix. Theorem 1 suggests that we can leverage the existing information directly for the first situation, and incrementally update the information for the second situation, only compute for the third situation.

Example. Consider the multi-label information system given in Table 1 at time t. Suppose at time $t+1$, four objects are immigrated, as shown in Table 2:

We can deduce the equivalence classes of $\Delta U^{t+1}/R_1^t$, $\Delta U^{t+1}/R_2^t$, $\Delta U^{t+1}/R_1^t$ as:

$\Delta U^{t+1}/R_1^t = \Delta U^{t+1}/\{c_1, c_3\} = \{\{x_{12}\}, \{x_{13}\}, \{x_{14}\}, \{x_{15}\}\}$,
$\Delta U^{t+1}/R_2^t = \Delta U^{t+1}/\{c_2, c_3\} = \{\{x_{12}\}, \{x_{13}, x_{15}\}, \{x_{14}\}\}$,
$\Delta U^{t+1}/R_3^t = \Delta U^{t+1}/\{c_1, c_2\} = \{\{x_{12}, x_{13}\}, \{x_{14}\}, \{x_{15}\}\}$.

By following Theorem 1, we can estimate the value of $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$, $\rho_{R_2^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_2)$, $\rho_{R_3^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_3)$.

Calculation of $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$: The result $U^t/\{c_1,c_3\} = \{E_1^t, E_2^t, E_3^t, E_4^t, E_5^t\} = \{\{x_1,x_3\},\{x_2\},\{x_4,x_5,x_7\},\{x_6,x_8,x_9,x_{11}\},\{x_{10}\}\}$ is known. Firstly, we examine the changes of existing equivalence classes.

$E_1^t(i.e.,\ \{x_1,x_3\})$: Among $\Delta U^{t+1}/R_1^t$, $\{x_{14}\}$ is with $(c_1,c_3) = (1,1)$. Additionally, $E_1^t \subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^t)$, where Y_1^t are the concept that instances are associated with label L_1 at time t. Besides, $x_{14} \in \underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^{t+1})$ The two conditions imply that the equivalence classes satisfies $E_1^t \subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^{t+1})$, and this contributes $\frac{2+1}{11+4} = \frac{3}{15}$ in $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$.

$E_2^t(i.e.,\ \{x_2\})$: Among $\Delta U^{t+1}/R_1^t$, no equivalence classes are with $(c_1,c_3) = (3,2)$. Additionally, $E_2^t \subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_2}(Y_2^t)$, where Y_2^t are the concept that instances are not associated with label L_1 at time t. The two conditions imply that the equivalence classes satisfies $E_2^t \subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_2}(Y_2^{t+1})$, and this contributes $\frac{1}{11+4} = \frac{1}{15}$ in $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$.

$E_3^t(i.e.,\ \{x_4,x_5,x_7\})$: Among $\Delta U^{t+1}/R_1^t$, the fourth equivalence class $\{x_{15}\}$ is with $(c_1,c_3) = (2,1)$. Additionally, $P(Y_1^t|E_3^t) = \frac{2}{3}$, $P(Y_2^t|E_3^t) = \frac{1}{3}$, where Y_1^t and Y_2^t are the concepts that instances are associated/not associated with label L_1 at time t respectively. By following Lemma 1, $P(Y_1^{t+1}|E_3^{t+1})$ is increased to $\frac{3}{4}$, which still fails to be subset of $\underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^{t+1})$ and subset of $\underline{\mathcal{R}}_{R_2^t}^{\alpha_1}(Y_2^{t+1})$. According to Lemma 2 and Theorem 1, this component has no contribution to $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$.

$E_4^t(i.e.,\ \{x_6,x_8,x_9,x_{11}\})$: Among $\Delta U^{t+1}/R_1^t$, the first equivalence class $\{x_{12}\}$ is with $(c_1,c_3) = (1,2)$. Additionally, $P(Y_1^t|E_4^t) = \frac{1}{2}$, $P(Y_2^t|E_4^t) = \frac{1}{2}$, where Y_1^t and Y_2^t are the concept that instances are associated/not associated with label L_1. This implies that $E_4 \not\subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^t)$. By following Lemma 1, $P(Y_1^{t+1}|E_4^{t+1})$ is increased to $\frac{3}{5}$, which still fails to be subset of $\underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^t)$ and subset of $\underline{\mathcal{R}}_{R_2^t}^{\alpha_1}(Y_2^t)$. According to Lemma 2 and Theorem 1, this component has no contribution to $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$.

$E_5^t(i.e.,\ \{x_{10}\})$: Among $\Delta U^{t+1}/R_1^t$, the second equivalence class $\{x_{13}\}$ is with $(c_1,c_3) = (3,1)$. Additionally, $P(Y_1^t|E_5^t) = \frac{1}{2}$ $E_2^t \subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^t)$, where Y_1^t are the concept that instances are not associated with label L_1 at time t. The two conditions imply that the equivalence classes satisfies $E_5^t \cup \{x_{10}\} \not\subset \underline{\mathcal{R}}_{R_1^t}^{\alpha_1}(Y_1^{t+1})$, and this component has no contribution to $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$.

Secondly, we evaluate the influence of new equivalence classes to $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1)$. However, no such equivalence classes exist.

Hence, $\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_1) = \frac{3}{15} + \frac{1}{15} = \frac{4}{15}$.

The calculation of $\rho_{R_2^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_2)$ and $\rho_{R_3^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_3)$ is analogous. $\rho_{R_2^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_2) = \frac{2}{5}$, $\rho_{R_3^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_3) = \frac{1}{3}$. □

For single-label case, the reduct remains unchanged if approximation quality is kept, i.e. $R_i^{t+1} = R_i^t$ if $\rho_{R_i^t}^{(\alpha,\beta)}(U^t/L_i) = \rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$. However, it does not invariably hold for multi-label case. The reasons are two-folds. Firstly, the mechanism of selective ensemble is vulnerable to the super-reduct, which signifies that the classification performance may be degenerated. Secondly, compared with finding a new reduction, a more compact representation on the same label may maintain the approximation degree with limited time complexity. To balance the effectiveness and efficiency, we consider the following strategies:

- $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i) = \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i)$: Find a reduct R_i^{t+1} with $R_i^{t+1} \subseteq R_i^t$ s.t.
$$\rho_{R_i^{t+1}}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i) = \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i),\ \rho_{R_i^{t+1}-\{b\}}^{(\alpha,\beta)}(U^t \cup \Delta U^{t+1}/L_i) \neq \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i).\forall b \in R_i^t$$
- $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i) \neq \rho_C^{(\alpha_1,\alpha_2)}(U^t/L_i)$: Find a reduct R_i^{t+1} with $R_i^{t+1} \not\subset R_i^t$ s.t.
$$\rho_{R_i^{t+1}}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i) = \rho_C^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i),\ \rho_{R_i^{t+1}-\{b\}}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i) \neq \rho_C^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i).\forall b \in R_i^t$$

It is worth mentioning that R_i^{t+1} may be the R_i^t if $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i) = \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i)$ holds. This means that the removal of attributes from R_i^t is not mandatory.

Example. Consider the multi-label information system given in Table 1 at time t. Suppose at time $t+1$, four objects are immigrated, as shown in Table 2.

For time t, we have:

$$\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^t/L_1) = \frac{4}{11},$$

$$\rho_{R_2^t}^{(\alpha_1,\alpha_2)}(U^t/L_2) = \frac{8}{11},$$

$$\rho_{R_3^t}^{(\alpha_1,\alpha_2)}(U^t/L_3) = \frac{5}{11}.$$

For time $t+1$, we have:

$$\rho_{R_1^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_1) = \frac{4}{15},$$
$$\rho_{R_2^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_2) = \frac{2}{5},$$
$$\rho_{R_3^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_3) = \frac{1}{3}.$$

Obviously, $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i) \neq \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t \cup \Delta U^{t+1}/L_i), i \in \{1,2,3\}$. Thus, the targeted reduct R_i^{t+1} satisfies $R_i^{t+1} \not\subset R_i^t$. □

4.1.2. Updating mechanism of granular correlation matrix

Recomputing granular structure correlation matrix (GM^t) is also time-consuming. Computing pair-wise granular structures which are definitely intersected does not introduce any benefit. It is worth mentioning that the non-empty judgment on reduction intersection is our interest, we can accelerate the computation from the perspective of set cardinality. We define a matrix with the same rank as granular structure correlation matrix called cardinality counting matrix (CCM). Let R_i^t and R_j^t be the reduct w.r.t. label i and label j at time t respectively, we have:

$$CCM^t = \{ccm_{i,j}^t\}_{|L|\times|L|}. \tag{21}$$

where

$$ccm_{i,j}^t = |R_i^t \cap R_j^t|.$$

Example. Consider the multi-label information system given in Table 1. Based on Eq. (21), the corresponding CCM^t is computed

as:

$$CCM^t = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \quad \square$$

Theorem 2. *Given a CCM^t, $\overline{\Delta R_i^t}$ and $\underline{\Delta R_i^t}$ represent the immigrated/emmigrated attributes of label-specific reduct w.r.t. label i at time t. Then for $\forall t > t_0$, we have*

$ccm_{i,j}^t > 0$.

if $ccm_{i,j}^{t-1} > 0 \wedge ccm_{i,j}^{t-1} + \left|\overline{\Delta R_i^t} \cap \overline{\Delta R_j^t}\right| - \left|\underline{\Delta R_i^t} \cup \underline{\Delta R_j^t}\right| > 0$.

The proof of Theorem 2 can be found in appendix.

Theorem 3. *Given a $MLIS^t$, the element $gm_{i,j}^t > 0$ in granular structure system GSS^t holds if the corresponding element $ccm_{i,j}^t$ in cardinality counting matrix CCM^t satisfies $ccm_{i,j}^t > 0$.*

The proof of Theorem 3 can be found in Appendix. Accordingly, we can simplify the problem of computing $gm_{i,j}^t$ as the determine of $ccm_{i,j}^t > 0$.

Example. Consider the multi-label information system given in Table 1. Suppose at time $t+1$, four objects are immigrated, as shown in Table 2. Then the variations w.r.t. R_1^t, R_2^t, and R_3^t are deduced as:

$\overline{\Delta R_1^t} = \{c_2\}$, $\underline{\Delta R_1^t} = \emptyset$;

$\overline{\Delta R_2^t} = \{c_1\}$, $\underline{\Delta R_2^t} = \emptyset$;

$\overline{\Delta R_3^t} = \{c_3\}$, $\underline{\Delta R_3^t} = \emptyset$.

Based on Theorem 2, we update $ccm_{i,j}^{t+1}$ as:

$ccm_{1,2}^{t+1}$: $ccm_{1,2}^t + \left|\overline{\Delta R_1^t} \cap \overline{\Delta R_2^t}\right| - \left|\underline{\Delta R_1^t} \cap \underline{\Delta R_2^t}\right| = 1+0-0 > 0$ therefore, $ccm_{1,2}^{t+1} > 0$ and $ccm_{1,2}^{t+1}$ is estimated as 1. This implies the features included in R_1^{t+1} are effective for the determine of L_2 at time $t+1$, and vice versa.

$ccm_{1,3}^{t+1}$: $ccm_{1,3}^t + \left|\overline{\Delta R_1^t} \cap \overline{\Delta R_3^t}\right| - \left|\underline{\Delta R_1^t} \cap \underline{\Delta R_3^t}\right| = 1+0-0 > 0$ therefore, $ccm_{1,3}^{t+1} > 0$ and $ccm_{1,3}^{t+1}$ is estimated as 1. This implies the features included in R_1^{t+1} are effective for the determine of L_3 at time $t+1$, and vice versa.

$ccm_{2,3}^{t+1}$: $ccm_{2,3}^t + \left|\overline{\Delta R_2^t} \cap \overline{\Delta R_3^t}\right| - \left|\underline{\Delta R_2^t} \cap \underline{\Delta R_3^t}\right| = 1+0-0 > 0$ therefore, $ccm_{2,3}^{t+1} > 0$ and $ccm_{2,3}^{t+1}$ is estimated as 1. This implies the features included in R_2^{t+1} are effective for the determine of L_3 at time $t+1$, and vice versa. $\square$

4.2. An incremental algorithm

Based on the incremental mechanisms of granular structure system, this subsection introduces an algorithm named Incremental Multiple Object with Three-way decisions and Selective ENsemble ("IMOTSEN", see Algorithm 1).

Time complexity for Algorithm 1 at time $t+1$ is analyzed as follows: Step 1 occupies $O(avg(R_i^t)|(U^t \cup \Delta U^{t+1})/avg(R_i^t) \parallel L|)$, Step 2 takes $\lambda_1 \times O((|C| - avg(R_i^t))^2|(U^t \cup \Delta U^{t+1})/C \| L|) + (1-\lambda_1) \times O(avg(R_i^t)^2|(U^t \cup \Delta U^{t+1})/avg(R_i^t) \| L|)$, where operator $avg(\cdot)$ represents the average length of a set and $0 \le \lambda_1 \le 1$. Step 3 costs $O(\lambda_2 \times avg(R_i^t)^2|L|^2 + |L|^2)$. $O((|C| - avg(R_i^t))^2|(U^t \cup \Delta U^{t+1})/C \| L|) \le O(|C|^2|(U^t \cup \Delta U^{t+1})/C \| L|)$ holds, and dominates the step 3, where $0 \le \lambda_2 \le 1$. Therefore, the overall complexity for Algorithm 1 is $O((|C| - avg(R_i^t))^2|(U^t \cup \Delta U^{t+1})/C \parallel L|)$.

Table 3
Description of datasets.

Dataset	# instances	# features	# labels	# cardinality
Genbase	662	1185	27	1.252
Medical	978	1449	45	1.245
Enron	1702	1001	53	3.39
Slashdot	3782	1079	22	1.18
LangLog	1460	1004	75	1.18
Bibtex	7395	1836	159	2.402

Algorithm 1: Incremental multiple-object three-way selective ensemble (IMOTSEN)

Input: (α_1, α_2), $GSS^t = (MLIS^t, IND(\cdot), R^t, GM^t, \frac{|R_i^t \cap R_j^t|}{|R_i^t \cup R_j^t|})$, $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i)\ \forall i, j \wedge i \neq j$, immigrated objects ΔU^{t+1}.

Output: $GSS^{t+1} = (MLIS^{t+1}, IND(\cdot), R^{t+1}, GM^{t+1}, \frac{|R_i^{t+1} \cap R_j^{t+1}|}{|R_i^{t+1} \cup R_j^{t+1}|})$

Step 1: Examination of $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$.
Step 1.1: Update $P([x]_{L_i} | [x]_{R_i^t})$ according to Lemma 2.
Step 1.2: Calculate $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$ according to Equation 20.
Step 2: Updating of granular structure R_i^t.
if $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i) = \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i)$, then
Step 2.1: Find $R_i^* \subseteq R_i^t$ s.t. $\rho_{R_i^*}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i) = \rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^t/L_i)$,
Step 2.2: Store the last appended attribute b and $RT_i^* = RT_i^* \cup \{b\}$.
Step 2.3: Loop from Step 2.1 to Step 2.2 until $\forall c \in R_i$. $\rho_{RT_i^* - \{c\}}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i) \neq \rho_{R_i^t}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$, and therefore get the granular structure $R_i^{t+1} = RT_i^*$.
else
Step 2.4: Compute $\rho_C^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$.
Step 2.5: Select $c = \arg\max_c \gamma_{R_i^t \cup \{c\}}((U^t \cup \Delta U^{t+1})/L_i)$, $c \in C - R_i^t$ $R_i^t = R_i^t \bigcup \{c\}$.
Step 2.6: Loop Step 2.5 until $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i) = \rho_C^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i)$.
Step 2.7: Store the last appended attribute b and $RT_i^* = RT_i^* \cup \{b\}$.
Step 2.8: Loop Step 2.7 until $\forall c \in RT_i^*$. $\rho_{RT_i^* - \{c\}}^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i) \neq \rho_C^{(\alpha_1,\alpha_2)}((U^t \cup \Delta U^{t+1})/L_i)$, and therefore get the granular structure $R_i^{t+1} = RT_i^*$.
Step 2.9: Loop from Step 2.1 to Step 2.8 $\forall L_i$.
Step 3: Updating of granular correlation matrix GM^t.
Step 3.1: Updating $gm_{i,j}^{t+1}$ according to Theorem 3.
Step 3.2: Loop until each L_i is visited to generate GM^{t+1}.

5. Experimental analysis

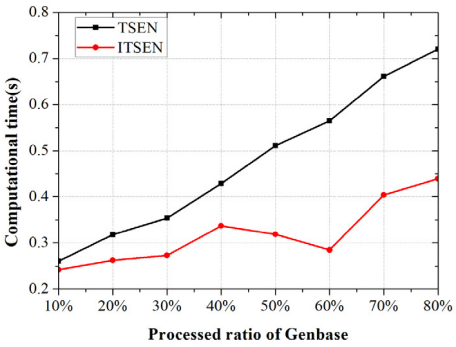
5.1. Datasets

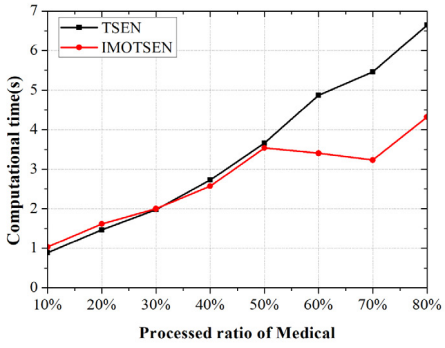
We conducted experiments on six multi-label benchmark datasets, the details of which are summarized in Table 3. To satisfy the equivalence relation requirement, attributes type of selected benchmark is nominal. The term "cardinality" is abbreviated for label cardinality representing the average labels count regarding instances. They can be downloaded from the websites of Mulan[5] [45] and Meka[6] [46].

5.2. Evaluation metrics

The primary goal behind the introduction of incremental mechanism is to overcome computation deficiency. The execution time for the construction of *GSS* in both non-incremental version

[5] http://mulan.sourceforge.net/datasets.html.
[6] http://meka.sourceforge.net/.





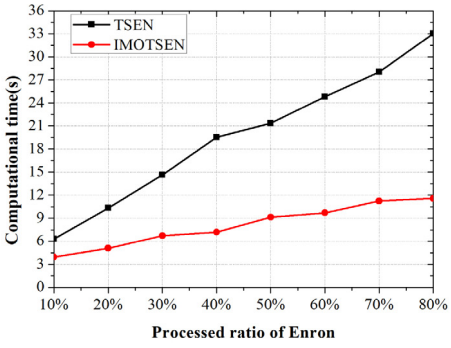
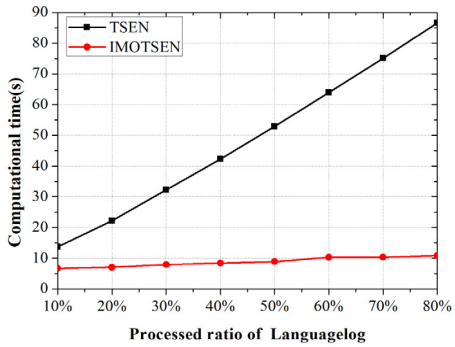
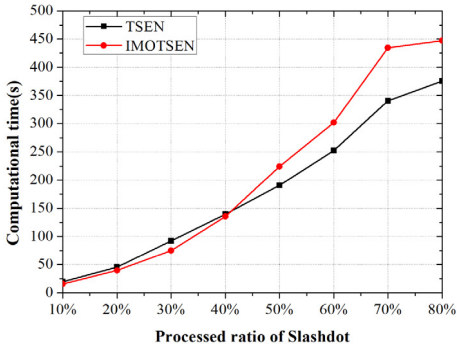


Fig. 2. Elapsed time between TSEN and IMOTSEN on six multi-label datasets with $(\alpha_1, \alpha_2) = (0.8, 1.0)$.





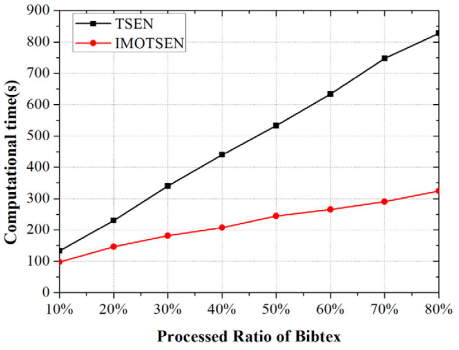


Fig. 2. (*continued*).

("TSEN") and incremental version ("IMOTSEN") over six benchmarks is included. All experiments are coded in Matlab 2017b and completed on a workstation with the following specification: Intel Core12 i7-6800K 3.40 GHz CPU, 64 GB of memory with 64-bit ubuntu 16.0.4 operation system.

Classification performance is another important aspect to evaluate the quality of renewed knowledge. The evaluation measures for multi-label classification is roughly classified into two taxonomies, i.e. *example-based* metrics and *label-based* metrics. The first category of metrics evaluates prediction performance on each example separately, and returns the mean value across the test set. The second category of metrics evaluates prediction performance on each label separately, and returns the mean value across all labels. We consider the following metrics *Label-based Precision* [47], *Label-based Recall* [47], *Example-based Precision* [48], *Example-based Recall* [48], *Hamming-Loss* [49], *Micro F1* [49]. Except for Hamming Loss, the larger the remaining metric values, the better the performance.

5.3. Experimental setting

For each benchmark in Table 2, 10% instances are randomly selected as basic datasets, and for the remaining instances, we divide them into nine subsets and regard each subset (i.e. 10% instances) as basic batch of immigrated objects. For a given threshold pair (α_1, α_2), we report eight round incremental tests for both time dimension and classification performance dimension (i.e. 10% object immigrations from current object assembles with the count ranges from 10% to 80% at a step of 10%). It is worth mentioning that all labels of batch objects are assumed to be available after each round of immigration.

The main goal of experiments is to testify whether "IMOTSEN" can generate comparable performance with satisfactory computational complexity. Thus, we examine three aspects of "IMOTSEN". The first experiment (see Section 5.4) evaluates the algorithm efficiency in terms of knowledge reconstruction/updating between "IMOTSEN" and "TSEN" with a predefined threshold pair $(\alpha_1, \alpha_2) = (0.8, 1.0)$. The second experiment (see Section 5.5) compares the classification performance between "IMOTSEN" and "TSEN" with the same threshold pair $(\alpha_1, \alpha_2) = (0.8, 1.0)$. The

Table 4
Average execution time for reconstruction of GSS between TSEN and IMOTSEN (unit: seconds).

Dataset	Alg-GSS	Size of processed ratio							
		10%	20%	30%	40%	50%	60%	70%	80%
Genbase	N-GS	0.225	0.283	0.320	0.391	0.474	0.527	0.625	0.683
	I-GS	0.219	0.241	0.251	0.318	0.300	0.257	0.384	0.418
	N-GM	0.036	0.035	0.034	0.037	0.037	0.037	0.036	0.036
	I-GM	0.022	0.021	0.022	0.019	0.019	0.027	0.020	0.021
Medical	N-GS	0.808	1.385	1.898	2.650	3.571	4.778	5.375	6.562
	I-GS	0.973	1.548	1.916	2.481	3.454	3.319	3.157	4.233
	N-GM	0.077	0.080	0.082	0.083	0.090	0.089	0.089	0.084
	I-GM	0.053	0.068	0.083	0.085	0.081	0.085	0.075	0.088
Enron	N-GS	6.178	10.19	14.51	19.39	21.24	24.70	27.89	32.90
	I-GS	3.806	4.943	6.584	7.049	8.988	9.557	11.10	11.42
	N-GM	0.135	0.140	0.140	0.139	0.142	0.142	0.141	0.132
	I-GM	0.126	0.126	0.127	0.126	0.130	0.125	0.129	0.131
Languagelog	N-GS	13.52	21.90	32.08	42.05	52.58	63.70	74.78	86.36
	I-GS	6.483	6.824	7.686	8.202	8.703	10.09	10.14	10.63
	N-GM	0.247	0.249	0.245	0.242	0.245	0.250	0.247	0.249
	I-GM	0.227	0.225	0.224	0.221	0.215	0.226	0.212	0.212
Slashdot	N-GS	19.46	45.80	92.29	139.3	190.5	252.3	340.4	375.6
	I-GS	15.49	39.48	74.99	135.6	224.0	301.9	434.0	447.3
	N-GM	0.027	0.029	0.028	0.027	0.024	0.025	0.025	0.025
	I-GM	0.039	0.041	0.043	0.038	0.034	0.035	0.038	0.035
Bibtex	N-GS	131.9	229.2	338.6	438.9	531.9	632.8	746.4	826.8
	I-GS	96.63	145.4	181.0	206.3	243.7	264.8	289.2	323.9
	N-GM	1.295	1.267	1.277	1.232	1.200	1.122	1.098	1.073
	I-GM	1.029	1.057	1.063	1.068	1.084	1.066	1.078	1.083

third experiment (see Section 5.6) explores the robustness of execution time of "IMOTSEN" under different threshold pairs $(\alpha_1, \alpha_2) \in \{(1.0, 1.0), (0.8, 1.0), (0.6, 1.0)\}$. All experiments are conducted with five-fold cross validation to conduct a fair comparison.

5.4. Performance comparison between TSEN and IMOTSEN: Algorithmic efficiency

The plots of computational efficiency between "TSEN" and "IMOTSEN" are displayed in Fig. 2. In each sub-figure, the x-axis is the count for instances with known labels, and the y-axis is the average value of computation time. Blocks marked with black lines are computation time of "TSEN" and circles marked with red lines are computation time of "IMOTSEN".

Fig. 2 illustrates the effectiveness of algorithm "TSEN" and algorithm "IMOTSEN". The time for reconstruction of GSS seems to be significantly compressed for dataset "Genbase", "Enron", "Languagelog" and "Bibtex". However, the result is a bit frustrated on dataset "Medical" and "Slashdot". To reveal the underlying reason, we decompose the reconstruction time for GSS as granular structure (abbreviated as GS) component and granular matrix(abbreviated as GM) component. For each dataset, we report four groups of time when objects are immigrated with 10% count. In what follows, "Alg-GSS" specifies the algorithm leveraged for updating granular structure system; "N-GS" signifies the time for reconstruction of granular structure when "TSEN" is employed; "T-GS" signifies the time for reconstruction of granular structure when "IMOTSEN" is employed; "N-GM" signifies the time for reconstruction of granular matrix when "TSEN" is employed; "T-GS" signifies the time for reconstruction of granular matrix when "IMOTSEN" is employed.

From Table 4, we can deduce that time reconstruction of granular structure is much longer than the reconstruction of granular matrix. The negligible time for generation of granular matrix implies that the overall time may still be saved if the time for updating of granular matrix is prolonged. This is consistent with the analysis in computation complexity. From Table 5 , we can observe that the time variations of renewing granular matrix are much smaller than finding granular structure. It implies that, compared to determine the shared attributes, the search for granular structure is more data-dependent.

As suggested in [50], the speed-up ratio can explain the relative superiority of efficiency between algorithms directly. The incremental speed-up ratio is defined as $\frac{T_{n-x}}{T_{i-x}}$, where $x \in \{GS, GM\}$, T_{n-x} represents execution time of the non-incremental algorithm (i.e. "TSEN") and T_{i-x} represents execution time of the incremental algorithm (i.e. "IMOTSEN"). Theoretically, "IMOTSEN" is assumed to be more efficient than "TSEN" if speed-up is larger than 1. The speed-up ratio that is larger than 1 is highlighted in bold face.

From Table 6 we can summarize that in most cases (77/96), "IMOTSEN" performs faster than "TSEN". For dataset "Genbase", "Enron", and "Languagelog", the speed-up (yields $1.027 - 8.124$) is rather acceptable for both GS and GM. For those speed-up times less than 1, the minimum is 0.651 (30% label known with 10% increase on "Slashdot"). Furthermore, the incremental mechanism is least effective on "Slashdot". It may due to the occurrence of concept drift in "Slashdot".

5.5. Performance comparison between TSEN and IMOTSEN: Classification accuracy

To visualize the closeness of classification performance, we draw spider web diagram to show the stability over six evaluation metrics. The referenced values for different metrics are affiliated in each graph to measure the accuracy/loss. For metric "Hamming Loss", the referenced value ranges from 0 to 0.2 with a step of 0.05. For metrics "Example-based Precision", "Example-based Recall", and "Micro F1" the referenced value ranges from 0 to 1 with a step of 0.2. For metrics "Label-based Precision" and "Label-based Recall", the referenced value ranges from 0 to 0.8 with a step of 0.2. Averaged performance on six evaluations generated from "TSEN" and "IMOTSEN" are circled in blue lines and red lines, respectively.

Table 5
Variance of time cost for reconstruction of GSS between TSEN and IMOTSEN.

Dataset	Alg-GSS	Size of processed ratio							
		10%	20%	30%	40%	50%	60%	70%	80%
Genbase	N-GS	0.016	0.042	0.049	0.072	0.063	0.059	0.034	0.043
	I-GS	0.026	0.025	0.022	0.059	0.058	0.059	0.049	0.119
	N-GM	0.003	0.002	0.003	0.004	0.003	0.002	0.001	0.001
	I-GM	0.003	0.004	0.004	0.002	0.003	0.005	0.005	0.001
Medical	N-GS	0.087	0.070	0.140	0.108	0.169	0.236	0.315	0.274
	I-GS	0.194	0.136	0.239	0.400	0.375	0.554	0.456	0.253
	N-GM	0.002	0.003	0.004	0.005	0.006	0.003	0.005	0.002
	I-GM	0.014	0.018	0.007	0.002	0.006	0.006	0.004	0.004
Enron	N-GS	0.673	0.927	2.027	2.602	4.351	3.891	3.892	5.023
	I-GS	0.196	0.349	0.499	0.826	1.159	0.575	0.732	0.839
	N-GM	0.009	0.009	0.009	0.008	0.008	0.011	0.011	0.006
	I-GM	0.003	0.004	0.003	0.005	0.003	0.004	0.002	0.003
Languagelog	N-GS	0.623	0.592	1.479	2.007	1.700	2.283	2.241	2.617
	I-GS	0.298	0.288	0.279	0.368	0.382	0.302	0.535	0.409
	N-GM	0.009	0.005	0.006	0.005	0.006	0.005	0.004	0.005
	I-GM	0.003	0.002	0.006	0.006	0.005	0.011	0.014	0.008
Slashdot	N-GS	1.724	5.397	10.01	17.44	21.89	62.26	46.42	48.04
	I-GS	2.283	6.809	5.478	18.70	34.93	44.41	31.68	163.491
	N-GM	0.003	0.003	0.002	0.001	0.003	0.001	0.002	0.001
	I-GM	0.002	0.004	0.003	0.004	0.001	0.002	0.004	0.002
Bibtex	N-GS	3.263	6.438	17.45	19.34	24.32	30.17	19.78	21.74
	I-GS	1.571	1.582	7.401	5.503	16.73	10.62	14.65	24.90
	N-GM	0.043	0.059	0.061	0.034	0.061	0.040	0.081	0.023
	I-GM	0.012	0.020	0.016	0.022	0.013	0.012	0.014	0.010

Table 6
The incremental speed-up for reconstruction of GSS between TSEN and IMOTSEN.

Dataset	GSS	Size of processed ratio							
		10%	20%	30%	40%	50%	60%	70%	80%
Genbase	GS	**1.027**	**1.174**	**1.275**	**1.230**	**1.580**	**2.051**	**1.628**	**1.634**
	GM	**1.636**	**1.667**	**1.545**	**1.947**	**1.947**	**1.370**	**1.800**	**1.714**
Medical	GS	0.830	0.895	0.991	**1.068**	**1.034**	**1.440**	**1.703**	**1.550**
	GM	**1.453**	**1.176**	0.988	0.976	**1.111**	**1.047**	**1.187**	0.955
Enron	GS	**1.623**	**2.062**	**2.204**	**2.751**	**2.363**	**2.584**	**2.513**	**2.881**
	GM	**1.071**	**1.111**	**1.102**	**1.103**	**1.092**	**1.136**	**1.093**	**1.008**
Languagelog	GS	**2.085**	**3.209**	**4.174**	**5.127**	**6.042**	**6.313**	**7.375**	**8.124**
	GM	**1.088**	**1.107**	**1.094**	**1.095**	**1.140**	**1.106**	**1.165**	**1.175**
Slashdot	GS	**1.256**	**1.160**	**1.231**	**1.027**	0.850	0.834	0.784	0.840
	GM	0.692	0.707	0.651	0.711	0.706	0.714	0.658	0.714
Bibtex	GS	**1.364**	**1.576**	**1.871**	**2.127**	**2.183**	**2.390**	**2.581**	**2.553**
	GM	**1.259**	**1.199**	**1.201**	**1.154**	**1.107**	**1.053**	**1.019**	0.991

From Fig. 3, we can observe that: (1) the classification performance deduced from *IMOTSEN* is data-dependent, which means the degree of instance inconsistency cannot be reduced significantly; (2) for all considered metrics, the shape of "TSEN" and "IMOTSEN" is very similar, which means that "IMOTSEN" obtains a comparable solution.

5.6. Robustness test for IMOTSEN with different threshold pairs

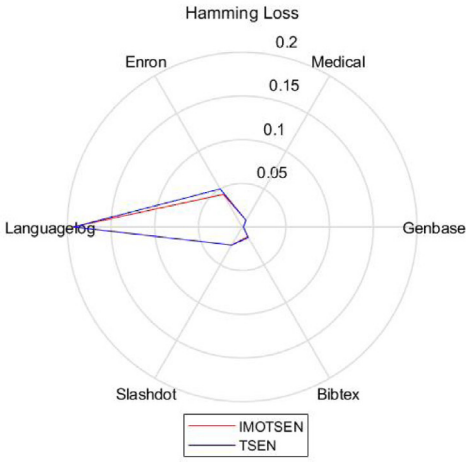
The accumulated computation time required for different datasets with the algorithm "IMOTSEN" is shown in Fig. 4. The x-axis corresponds to the value of threshold pair (α_1, α_2), and the y-axis corresponds to the computation time.

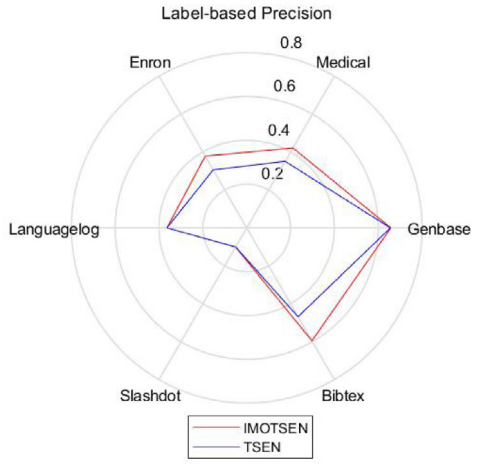
From Fig. 4, we can conclude that fluctuations of time on each dataset is very limited with different thresholds of (α_1, α_2). It further validates that the computation time depends mostly on data scale, which is also consistent with the analysis of algorithmic complexity.

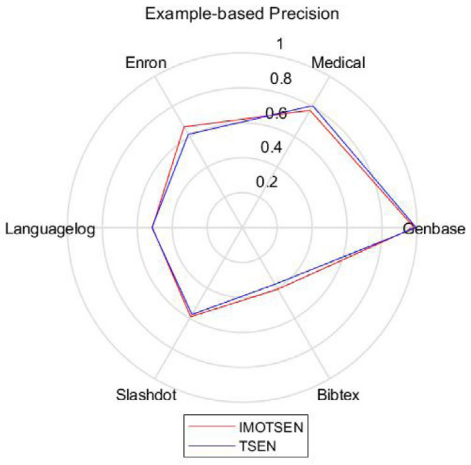
6. Discussion

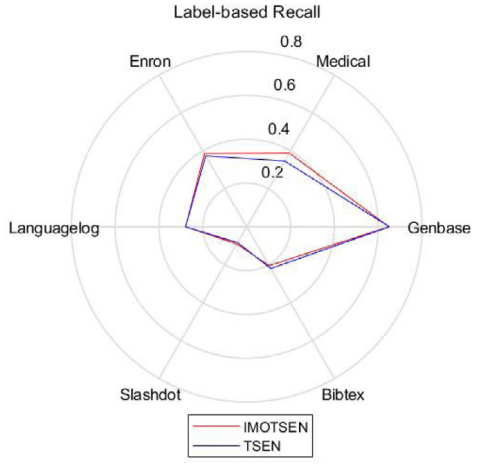
Combining Sections 5.4–5.6, we can conclude that, without losing too much classification accuracy, algorithm "IMOTSEN" can accelerate knowledge updating. In this section, we attempt to discuss some facets of "IMOTSEN".

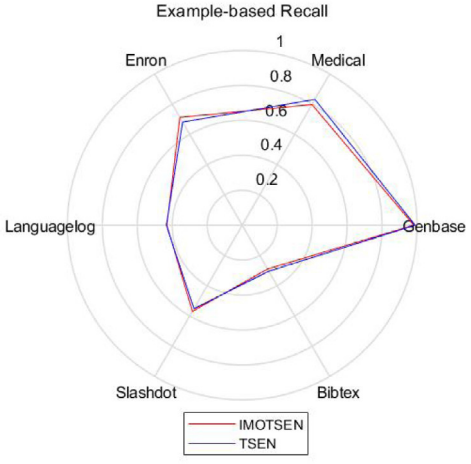
The first issue concerns the computation time. The experimental results seem to be contradicted by the complexity analysis. For updating of granular structure, we believe that in some cases, the strategy of taking a granular structure at time $t-1$ as the starting point of t may incur erroneous heuristic information. This means more attributes may be combined first to satisfy the ρ-preserving requirement. Additionally, for cases that ρ-preserving holds in both $t-1$ and t, the incremental mechanism seems to be more prudent in confirming the components of the granular structure. This operation can also introduce computation. It is also interesting to find that the acceleration of granular structure is not equivalent to the acceleration of granular matrix (see speed-up of "Slashdot" in Table 5), and vice versa (see speed-up of











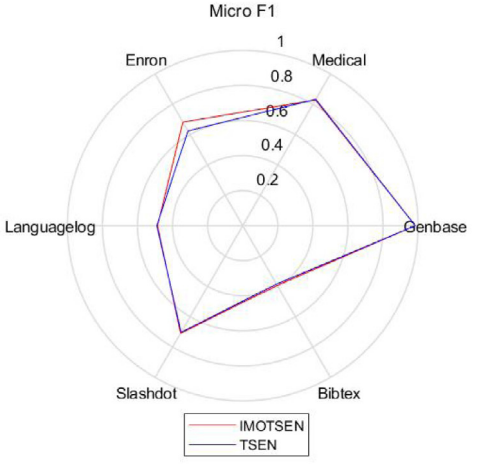


Fig. 3. (*continued*).

Fig. 3. Classification performance between TSEN and IMOTSEN on six multi-label datasets with $(\alpha_1, \alpha_2) = (0.8, 1.0)$.

"Medical" in Table 5). The reasons are two folds: (1) the limited variations of reductions may influence the shared attributes, leading to more non-positive elements in *CCM*. The acceleration

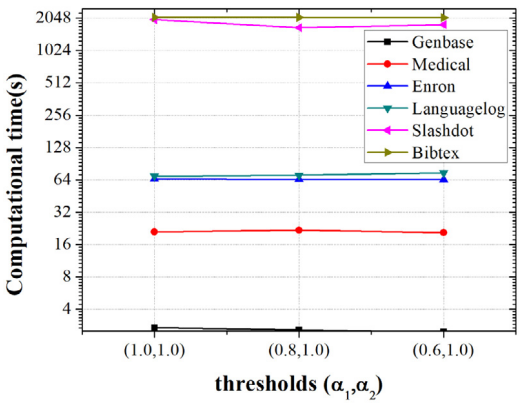


Fig. 4. Execution time for updating GSS with IMOTSEN on multi-label datasets.

fails if the proportion of non-positive elements in *CCM* is large enough, which means the computation of *CCM* is extra operation in "TSEN". (2) the positive elements of *CCM* do not signify that the granular structure can be preserved, which means the variations of granular structure have limited influence on the result of granular matrix.

Another issue is related to the strategy of updating granular structure when ρ-preserving strategy holds (see Section 4.1.1). Originally, we intend to keep the reducts unchanged when ρ-preserving strategy holds. However, we find that the classification performance on "Genbase" and "Medical" degenerate dramatically as more labels known (for "Genbase", the performance on "Micro F1" reduced from 0.93 to 0.68, and for "Medical", the performance on "Micro F1" reduced from 0.81 to 0.53). It is reasonable since the model capacity, a metric to evaluate the fitness of hypothesis space, is competent for a limited number of instances with large size of features and labels. As more instances are immigrated, the candidate solutions are gradually focused. We believe that an incremental mechanism does not make any sense if fast knowledge updating is obtained at the expense of deteriorated quality. The change of updating strategy on granular structure implies the value of controlled variations.

7. Conclusions

Labeling is a time-consuming work in practical applications, and the cost in both time and money to obtain high-quality label in multi-label data is exponential. To cope with multi-label effectively, we propose a novel granular structure system from the perspective of label-specific. Motivated by the knowledge updating work in single-label environment, we have extended three-way selective algorithm ("TSEN") from a static environment to dynamical environment ("IMOTSEN"). The introduction of incremental mechanism on granular structure system can significantly improve the computation efficiency with the maintenance of classification performance. In the future work, we will consider other scenarios of multi-label incremental learning. Real applications will also be examined to demonstrate the feasibility of proposed algorithm.

Acknowledgments

Authors would like to thank the anonymous reviewers for their constructive comments and valuable suggestions. This work is supported by the National Science Foundation of China (Grant No. 61976158, 61673301, 61763031, 61563016, 61573255, 61906137), National Key R&D Program of China (Grant No. 213), and Major Project of Ministry of Public Security, China (Grant No. 20170004). The first author thanks all the contributions from his supervisors and colleagues, particularly for the great support from lover Tianna Zhao. Will you spend the rest of life with me?

Appendix. Proof of Lemma 1

Proof.

$$P(Y_j^{t+1}|B_i^{t+1}) = \frac{|Y_j^{t+1} \cap B_i^{t+1}|}{|B_i^{t+1}|} = \frac{|(Y_j^t \cup \Delta E_{Y_j}^{t+1}) \cap (B_i^t \cup \Delta E_{B_i}^{t+1})|}{|B_i^t \cup \Delta E_{B_i}^{t+1}|} = \frac{|Y_j^t \cap B_i^t| + |\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}|}{|B_i^t| + |\Delta E_{B_i}^{t+1}|}.$$

(1) if $P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) \geq P(Y_j^t|B_i^t)$, we have

$$P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) = \frac{|\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}|}{|\Delta E_{B_i}^{t+1}|} \geq \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} = P(Y_j^t|B_i^t).$$

which means $P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) = \frac{|Y_j^t \cap B_i^t|+N}{|B_i^t|}$, for some $N \geq 0$. Suppose $|\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}| = k \times (|Y_j^t \cap B_i^t| + N)$, where $k > 0$, then we have $|\Delta E_{B_i}^{t+1}| = k \times (|B_i^t|)$. Hence, we can deduce that

$$P(Y_j^{t+1}|B_i^{t+1}) = \frac{|Y_j^t \cap B_i^t| + |\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}|}{|B_i^t| + |\Delta E_{B_i}^{t+1}|} = \frac{|Y_j^t \cap B_i^t| + k \times (|Y_j^t \cap B_i^t| + N)}{|B_i^t| + k \times (|B_i^t|)} = \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} + \frac{kN}{(1+k) \times |B_i^t|} \geq \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} = P(Y_j^t|B_i^t).$$

(2) if $P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) < P(Y_j^t|B_i^t)$, then we have

$$P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) = \frac{|\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}|}{|\Delta E_{B_i}^{t+1}|} < \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} = P(Y_j^t|B_i^t).$$

which means $P(\Delta E_{Y_j}^{t+1}|\Delta E_{B_i}^{t+1}) = \frac{|Y_j^t \cap B_i^t|-N}{|B_i^t|}$, for some $N > 0$. Suppose $|\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}| = k \times (|Y_j^t \cap B_i^t| - N)$, where $k > 0$, then we have $|\Delta E_{B_i}^{t+1}| = k \times (|B_i^t|)$. Hence, we can deduce that

$$P(Y_j^{t+1}|B_i^{t+1}) = \frac{|Y_j^t \cap B_i^t| + |\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}|}{|B_i^t| + |\Delta E_{B_i}^{t+1}|} = \frac{|Y_j^t \cap B_i^t| + k \times (|Y_j^t \cap B_i^t| - N)}{|B_i^t| + k \times (|B_i^t|)} = \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} - \frac{kN}{(1+k) \times |B_i^t|} < \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} = P(Y_j^t|B_i^t).$$

(3) if $\Delta E_{B_i}^{t+1} = \emptyset$, then we have

$$P(Y_j^{t+1}|B_i^{t+1}) = \frac{|Y_j^t \cap B_i^t| + |\Delta E_{Y_j}^{t+1} \cap \Delta E_{B_i}^{t+1}|}{|B_i^t| + |\Delta E_{B_i}^{t+1}|} = \frac{|Y_j^t \cap B_i^t|}{|B_i^t|} = P(Y_j^t|B_i^t). \quad \square$$

Proof of Lemma 2

Proof.

$$\underline{\mathcal{R}}_B(Y_j^{t+1}) = POS_B^{\alpha_j}(Y_j^{t+1}) = \{x \in U^t \cup \Delta U^{t+1} | P(Y_j^{t+1}|[x]_B) \geq \alpha_j\}.$$

(1) Given $P(Y_j^t|B_i^t) \geq \alpha_j$, then for all B_i which satisfies $P(Y_l^{t+1}|B_i^{t+1}) \geq \alpha_k$, we can deduce that $B_i^{t+1} \subseteq POS_B^{\alpha_j}(Y_j^{t+1})$ and $B_i^t \subseteq POS_B^{\alpha_j}(Y_j^t)$. Since $\Delta E_{B_i}^{t+1} \in \Delta U^{t+1}/B$ and $B_i^{t+1} = B_i^t \cup \Delta E_{B_i}^{t+1}$, we have $\Delta E_{B_i}^{t+1} \subseteq POS_B^{\alpha_j}(Y_j^{t+1})$. Thus we should append all these $\Delta E_{B_i}^{t+1}$ as a component of $\underline{\mathcal{R}}_B(Y_j^{t+1})$, i.e.

$$\underline{\mathcal{R}}_B(Y_j^{t+1}) = \underline{\mathcal{R}}_B(Y_j^t) \cup \Delta E_{B_i}^{t+1}.$$

(2) Given $P(Y_j^t|B_i^t) \geq \alpha_j$, then for all B_i which satisfies $P(Y_j^{t+1}|B_i^{t+1}) < \alpha_j$, we can deduce that $B_i^{t+1} \notin POS_B^{\alpha_j}(Y_j^{t+1})$ and $B_i^t \subseteq POS_B^{\alpha_j}(Y_j^t)$. Since $\Delta E_{B_i}^{t+1} \in \Delta U^{t+1}/B$ and $B_i^{t+1} = B_i^t \cup \Delta E_{B_i}^{t+1}$, we have $B_i^t \notin POS_B^{\alpha_j}(Y_j^{t+1})$. Thus, for $\underline{\mathcal{R}}_B(Y_j^{t+1})$ we should delete all these B_i^t from $POS_B^{\alpha_j}(Y_j^t)$, i.e.

$$\underline{\mathcal{R}}_B(Y_j^{t+1}) = \underline{\mathcal{R}}_B(Y_j^t) - B_i^t.$$

(3) Given $P(Y_j^t|B_i^t) < \alpha_j$, then for all B_i which satisfies $P(Y_j^{t+1}|B_i^{t+1}) \geq \alpha_k$, we can deduce that $B_i^{t+1} \subseteq POS_B^{\alpha_j}(Y_j^{t+1})$ and $B_i^t \notin POS_B^{\alpha_j}(Y_j^t)$. Since $\Delta E_{B_i}^{t+1} \in \Delta U^{t+1}/B$ and $B_i^{t+1} = B_i^t \cup \Delta E_{B_i}^{t+1}$, we have $B_i^t \subseteq POS_B^{\alpha_j}(Y_j^{t+1})$ and $\Delta E_{B_i}^{t+1} \subseteq POS_B^{\alpha_j}(Y_j^{t+1})$. Thus, for $\underline{\mathcal{R}}_B(Y_j^{t+1})$ we should append all these B_i^t and $\Delta E_{B_i}^{t+1}$ as a component of $\underline{\mathcal{R}}_B(Y_j^{t+1})$, i.e.

$$\underline{\mathcal{R}}_B(Y_j^{t+1}) = \underline{\mathcal{R}}_B(Y_j^t) \cup B_i^t \cup \Delta E_{B_i}^{t+1}.$$

(4) Given $\Delta E_{B_i}^{t+1} = \emptyset$, then for all B_i, we can deduce that $B_i^{t+1} = B_i^t$, which implies $P(Y_j^{t+1}|[x]_B) = P(Y_j^t|[x]_B)$, i.e.

$$\underline{\mathcal{R}}_B(Y_j^{t+1}) = \underline{\mathcal{R}}_B(Y_j^t).$$

(5) Given $P(Y_j^t|B_i^t) < \alpha_j$, then for all B_i which satisfies $P(Y_j^{t+1}|B_i^{t+1}) < \alpha_j$, we can deduce that $B_i^{t+1} \notin POS_B^{\alpha_j}(Y_j^{t+1})$ and $B_i^t \notin POS_B^{\alpha_j}(Y_j^t)$. Since $\Delta E_{B_i}^{t+1} \in \Delta U^{t+1}/B$ and $B_i^{t+1} = B_i^t \cup \Delta E_{B_i}^{t+1}$, we have $B_i^t \notin POS_B^{\alpha_j}(Y_j^{t+1})$, which implies addition of $\Delta E_{B_i}^{t+1}$ has no influence on $POS_B^{\alpha_j}(Y_j^{t+1})$, i.e.

$$\underline{\mathcal{R}}_B(Y_j^{t+1}) = \underline{\mathcal{R}}_B(Y_j^t). \quad \square$$

Proof of Theorem 1

Proof. According to Eq. (13), we can deduce that $\rho_{R_i^{t+1}}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i) = \frac{|\underline{\mathcal{R}}_{R_i^{t+1}}^{\alpha_1}(Y_1^{t+1})| + |\underline{\mathcal{R}}_{R_i^{t+1}}^{\alpha_2}(Y_2^{t+1})|}{|U^{t+1}|}$ Since $U^{t+1} = U^t \cup \Delta U^{t+1}$, we have $\left|U^{t+1}\right| = \left|U^t \cup \Delta U^{t+1}\right| = \left|U^t\right| + \left|\Delta U^{t+1}\right|$. Meanwhile, the condition ΔU^{t+1} is composed of ΔU_1^{t+1} and ΔU_2^{t+1} implies that equivalence classes from either $\Delta U_1^{t+1}/R_i^t$ or $\Delta U_2^{t+1}/R_i^t$ have the result of empty set with the intersection of U^t/R_i^{t+1}. Given $R_i^{t+1} \leftarrow R_i^t$, we will demonstrate the three components subsequently.

(1) For the first q equivalence classes (i.e. $E_r^t, r \in \{1, 2, \ldots, q\}$), we have $\Delta E_r^{t+1} = \emptyset$. Furthermore, for such E_r^t, they will be recognized as component of lower approximation with regard to Y_k at time $t+1$ if it satisfies condition $E_r^t \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_k}(Y_j^{t+1})$. Accordingly, we denote $\frac{\sum_{r=1}^{j}\left|\bigcup\{E_r^t | E_r^t \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})\}\right|}{|U^t|+|\Delta U^{t+1}|}$ as the first component of $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i)$.

(2) For the middle s equivalence classes (i.e. $E_r^t, r \in \{q+1, q+2 \cdots, q+s\}$), the condition $E_r^t \cup \Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})$ implies that both $E_r^t \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})$ and $\Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})$. From Definition 3, we have $P(Y_j^{t+1}|E_i^t) \geq \alpha_k$ and $P(Y_j^{t+1}|\Delta E_i^{t+1}) \geq \alpha_j$. Obviously, we have $P(Y_j^t|E_i^t) \geq \alpha_k$. Based on Lemma 2, we have $\underline{\mathcal{R}}_{R_i^t}(Y_k^{t+1}) = \underline{\mathcal{R}}_{R_i^t}(Y_j^t) \cup \Delta E_{R_i^t}^{t+1}$. Accordingly, we denote $\frac{\sum_{r=q+1}^{q+s}\left|\bigcup\{E_r^t \cup \Delta E_r^{t+1} | E_r^t \cup \Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})\}\right|}{|U^t|+|\Delta U^{t+1}|}$ as the second component of $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i)$.

(3) For the last v equivalence classes (i.e. $E_r^t, r \in \{q+s+1, q+s+2 \cdots, q+s+v\}$), they have empty intersections with regard to all $E_r^t \in U^t/R_i^{t+1}$. the condition $\Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})$ implies that $\Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})$. From Definition 3, we have $P(Y_j^{t+1}|\Delta E_i^{t+1}) \geq \alpha_k$. Based on Lemma 2, we have $\underline{\mathcal{R}}_{R_i^t}(Y_j^{t+1}) = \underline{\mathcal{R}}_{R_i^t}(Y_j^t) \cup \Delta E_{R_i^t}^{t+1}$. Accordingly, we denote $\frac{\sum_{r=q+s+1}^{q+s+v}\left|\bigcup\{\Delta E_r^{t+1} | \Delta E_r^{t+1} \subseteq \underline{\mathcal{R}}_{R_i^t}^{\alpha_j}(Y_j^{t+1})\}\right|}{|U^t|+|\Delta U^{t+1}|}$ as the third component of $\rho_{R_i^t}^{(\alpha_1,\alpha_2)}(U^{t+1}/L_i)$. $\square$

Proof of Theorem 2

Proof. $ccm_{i,j}^{t-1} > 0$ suggests that $R_i^{t-1} \cap R_j^{t-1} \neq \emptyset$. $ccm_{i,j}^{t-1} + \left|\overline{\Delta R_i^t} \cap \overline{\Delta R_j^t}\right| - \left|\underline{\Delta R_i^t} \cup \underline{\Delta R_j^t}\right| > 0$ measures the lower bound of variations of intersecting attributes to removal attributes. Here lower bound means that elements in $\overline{\Delta R_i^t} \cap \overline{\Delta R_j^t}$ are definitely a subset of R_i^t and R_j^t whereas the attributes in $\underline{\Delta R_i^t} \cup \underline{\Delta R_j^t}$ may not reduce the size of the intersection of R_i^{t-1} and R_j^{t-1}. Since $\underline{\Delta R_i^t} \cap \overline{\Delta R_i^t} = \emptyset$ and $\underline{\Delta R_j^t} \cap \overline{\Delta R_j^t} = \emptyset$ hold, varied attributes will not be repetitively counted. Hence we complete the proof. $\square$

Proof of Theorem 3

Proof. Without losing generality, we consider the relation between $ccm_{i,j}^t$ and $gm_{i,j}^t$. The positive of $ccm_{i,j}^t$, which implies $\left|R_i^t \cap R_j^t\right| > 0$. According to Definition 7, $gm_{i,j}^t > 0$ holds. This completes the proof. $\square$

References

[1] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: Perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[2] Z. Pawlak, Rough Sets Int. J. Inf. Comput. Sci. (1982) 341–356.
[3] Q.A. Al-Radaideh, G.Y. Al-Qudah, Application of rough set-based feature selection for arabic sentiment analysis, Cogn. Comput. 9 (1) (2017) 1–10.
[4] S. Kumar, P. Kumar, Upper approximation based privacy preserving in online social networks, Expert Syst. Appl. 88 (2017) 276–289.
[5] L. Zhao, L. Shang, Y. Gao, Y.B. Yang, Video behavior analysis using topic models and rough sets, IEEE Comput. Intell. Mag. 8 (1) (2013) 56–67.
[6] X. Yang, T. Li, D. Liu, H. Chen, C. Luo, A unified framework of dynamic three-way probabilistic rough sets, Inform. Sci. 420 (2017) 126–147.

[7] D. Leite, R.M. Palhares, V.C.S. Campos, F. Gomide, Evolving granular Fuzzy model-based control of nonlinear dynamic systems, IEEE Trans. Fuzzy Syst. 23 (4) (2015) 923–938.

[8] J.H. Yu, W.H. Xu, Incremental computing approximations with the dynamic object set in interval-valued ordered information system, Fund. Inform. 142 (1–4) (2015) 373–397.

[9] C. Luo, T.R. Li, Z. Yi, H. Fujita, Matrix approach to decision-theoretic rough sets for evolving data, Knowl.-Based Syst. 99 (2016) 123–134.

[10] R.T. Das, K.A. Kai, Q. Chai, ieRSPOP: A novel incremental rough set-based pseudo outer-product with ensemble learning, Appl. Soft Comput. 46 (2016) 170–186.

[11] J.F. Xu, D.Q. Miao, Y.J. Zhang, Z.F. Zhang, A three-way decisions model with probabilistic rough sets for stream computing, Internat. J. Approx. Reason. 88 (2017) 1–22.

[12] Y. Yang, D.G. Chen, H. Wang, E.C.C. Tsang, D.L. Zhang, Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving, Fuzzy Sets Syst. 312 (2016) 66–86.

[13] G.M. Lang, D.Q. Miao, M.J. Cai, Z.F. Zhang, Incremental approaches for updating reducts in dynamic covering information systems, Knowl.-Based Syst. 134 (2017) 85–104.

[14] Y.G. Jing, T.R. Li, H. Fujita, Z. Yu, B. Wang, An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view, Inform. Sci. 411 (2017) 23–38.

[15] X. Xie, X. Qin, A novel incremental attribute reduction approach for dynamic incomplete decision systems, Internat. J. Approx. Reason. 93 (2018) 443–462.

[16] E. Gibaja, S. Ventura, A tutorial on multilabel learning, ACM Comput. Surv. 47 (3) (2015) 1–38.

[17] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.

[18] C. Lin, W.Q. Chen, C. Qiu, Y.F. Wu, S. Krishnan, Q. Zou, LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy, Neurocomputing 123 (2014) 424–435.

[19] J. Fuernkranz, E. Hullermeier, E.L. Menica, K. Brinker, Multilabel classification via calibrated label ranking, Mach. Learn. 73 (2008) 133–153.

[20] M.L. Zhang, L. Wu, LIFT: Multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 107–120.

[21] J.H. Xu, J.L. Liu, J. Yin, C.Y. Sun, A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously, Knowl.-Based Syst. 98 (2016) 172–184.

[22] Multi-label learning with label-specific features by resolving label correlations, Knowl.-Based Syst. 159 (2018) 148–157.

[23] L.Y. Song, J. Liu, B.Y. Qian, M.X. Sun, K. Yang, M. Sun, S. Abbas, A deep multi-modal cnn for multi-instance multi-label image classification, IEEE Trans. Image Process. 27 (2018) 6025–6038.

[24] Z.W. Shi, Y.M. Wen, Y. Xue, G.Y. Cai, Efficient class incremental learning for multi-label classification of evolving data streams, in: International Joint Conference on Neural Networks, 2014, pp. 2093–2099.

[25] Y.J. Lin, Q.H. Hu, J. Zhang, X.D. Wu, Multi-label feature selection with streaming labels, Inform. Sci. 372 (2016) 256–275.

[26] J.H. Liu, Y.J. Lin, S.X. Wu, C.X. Wang, Online multi-label group feature selection, Knowl.-Based Syst. 143 (2018) 42–57.

[27] Y. Zhu, K.M. Ting, Z.H. Zhou, Multi-label learning with emerging new labels, IEEE Trans. Knowl. Data Eng. 30 (10) (2018) 1902–1914.

[28] T.T. Nguyen, T. Nguyen, A. Luong, Q.V.H. Nguyen, A.W.C. Liew, B. Stantic, Multi-label classification via label correlation and first order feature dependance in a data stream, Pattern Recognit. 90 (2019) 35–51.

[29] J. Duan, Q.H. Hu, Z.L. J., Y.H. Qian, D.Y. Li, Feature selection for multi-label classification based on neighborhood rough sets, J. Comput. Res. Dev. 52 (2015) 56–65 (in Chinese).

[30] S.P. Xu, X.B. Yang, H.L. Yu, D.J. Yu, J.Y. Yang, E.C.C. Tsang, Multi-label learning with label-specific feature reduction, Knowl.-Based Syst. 104 (2016) 52–61.

[31] Y.J. Lin, Y.W. Li, C.X. Wang, J.K. Chen, Attribute reduction for multi-label learning with fuzzy rough set, Knowl.-Based Syst. 152 (15) (2018) 51–61.

[32] Y.W. Li, Y.J. Lin, J.H. Liu, W. Weng, Z. Shi, S.X. Wu, Feature selection for multi-label learning based on kernelizaed fuzzy rough sets, Neurocomputing 318 (2018) 271–286.

[33] H. Li, D.Y. Li, Y.H. Zhai, S.G. Wang, J. Zhang, A novel attribute reduction approach for multi-label data based on rough set theory, Inform. Sci. 367–368 (2016) 827–847.

[34] C. Cao, Y.Y. Yao, Actionable strategies in three-way decisions, Knowl.-Based Syst. 133 (2017) 141–155.

[35] Y.Y. Yao, Three-way decision: An interpretation of rules in rough set theory, in: International Conference on Rough Sets and Knowledge Technology, 2009, pp. 642–649.

[36] Y.Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inform. Sci. 181 (6) (2011) 1080–1096.

[37] Y.Y. Yao, Three-way decision and granular computing, Internat. J. Approx. Reason. 103 (2018) 107–123.

[38] G.M. Lang, M.D. Q., H. Fujita, Three-way group conflict analysis based on pythagorean fuzzy set theory, IEEE Trans. Fuzzy Syst. (2019) http://dx.doi.org/10.1109/TFUZZ.2019.2908123.

[39] W.W. Li, Z.Q. Huang, Q. Li, Three-way decisions based software defect prediction, Knowl.-Based Syst. 91 (2016) 263–274.

[40] G. Napoles, R. Falcon, E. Papageorgious, R. Bello, Rough cognitive ensembles, Internat. J. Approx. Reason. 85 (2017) 79–96.

[41] Y.J. Zhang, D.Q. Miao, Z.F. Zhang, J.F. Xu, S. Luo, A three-way selective ensemble model for multi-label classification, Internat. J. Approx. Reason. 103 (2018) 394–413.

[42] Y.Y. Yao, Y. Zhao, J. Wang, On reduct construction algorithms, in: Rough Sets & Knowledge Technology Proceedings, vol. 4062, 2006, pp. 100–117.

[43] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1992.

[44] Y.Y. Yao, S.K. Wong, A decision theoretic framework for approximating concepts, Int. J. Man-Mach. Stud. 37 (6) (1992) 793–809.

[45] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, I. Vlahavas, MULAN: A java library for multi-label learning, J. Mach. Learn. Res. 12 (7) (2012) 2411–2414.

[46] J. Read, P. Reutemann, B. Pfahringer, G. Holmes, Meka: a multi-label/multi-target extension to weka, J. Mach. Learn. Res. 17 (1) (2016) 667–671.

[47] G. Tsoumakas, I. Vlahavas, Random k-Labelsets: An ensemble method for multilabel classification, in: European Conference on Machine Learning, ECML, 2007, pp. 406–417.

[48] S. Godbole, S. Sarawagi, Discriminative methods for multi-labeled classification, Lecture Notes in Comput. Sci. 3056 (2004) 22–30.

[49] R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Mach. Learn. 37 (3) (1999) 297–336.

[50] C. Luo, T.R. Li, C.H. M., H. Fujita, Efficient updating of probabilistic approximations with incremental objects, Knowl.-Based Syst. 109 (1) (2016) 71–83.