Knowledge-Based Systems 252 (2022) 109472

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-attention concept-cognitive learning model: A perspective from conceptual clustering

Weihua Xu *, Yaoqi Chen

College of Artificial Intelligence, Southwest University, Chongqing, 400715, PR China

ARTICLE INFO

Article history:
Received 22 February 2022
Received in revised form 12 July 2022
Accepted 13 July 2022
Available online 19 July 2022

Keywords:
Concept-cognitive learning
Concept lattices
Conceptual clustering
Graph attention
Granular computing

ABSTRACT

Concept-cognitive learning (CCL), as a cognitive process, is an emerging field of simulating the human brain to learn concepts in the formal context. Simultaneously, attention is a core property of all perceptual and cognitive operations. Nevertheless, no current existing CCL models and conceptual clustering methods consider the impact of attention. In light of these observations, in this article, we present a novel concept learning method, called the multi-attention concept-cognitive learning model (MA-CLM), to address the issue by exploiting graph attention and the graph structure of the concept space. This model is deployed toward the goal of conceptual cognitive more reasonable: generate pseudo-concept with higher expected utility while taking into consideration making classification tasks more efficient. Specifically, a conceptual attention space is learned for each decision class via attribute attention. Furthermore, a new concept clustering and concept generation method based on graph attention was proposed based on the conceptual attention space. Comparative studies with S2CL$^{\alpha}$ over a total of nine UCI data sets validate the effectiveness and efficiency of concept clustering based on graph attention in concept-cognitive learning. In addition, we also performed a comparative evaluation of MA-CLM against several classical classification algorithms to demonstrate the excellent properties in classification tasks. Finally, the model is validated by concept generation on the handwritten numeral dataset MNIST.

1. Introduction

Cognitive computing is one of the core technical fields of cognitive science and an important part of artificial intelligence. More specifically, it is a computer system that simulates the cognitive process of the human brain. Concept-cognitive learning (CCL) is an emerging field that simulates the human cognitive process of concept learning, which provides a novel and effective method for knowledge discovery problems such as classification tasks [1–6], image annotation tasks [7] and rule extraction [8–10]. Numerous efforts have since continued to push the development of conceptual learning models at the theoretical level and application levels.

Concept-cognitive learning, to a large extent, depends on the structures of concept and the target concepts [11]. Accordingly, a large amount of concepts such as formal concept [12], fuzzy concept [13], object-oriented concept [14], attribute oriented concept [15], three-way concept [16], approximate concept [17] and AFS [18] concept have been proposed respectively. With the rapid growth of data size, if the concept sets do not have a strict lattice structure like the concept lattice, it may be able to obtain an effective and efficient learning algorithm [11]. In addition, granular computing theory believes that data can often be divided into different granular to meet different needs. Therefore, granular concepts [19], that is, concept space is introduced into CCL. In recent years, Shi [2,5] and Mi [3,4,20] propose several CCL models based on rule formal decision context and concept space to obtain conceptual generalization capability and deal with classification tasks. We show the development stages of the concept-cognitive learning in Fig. 1.

Conceptual clustering [31] is not only one of the essential methods in inductive learning (also known as concept learning [32]), but also a crucial element of the product development process [33]. Conceptual clustering focuses on dividing concepts into different categories, and then using new concepts with more vital generalization capabilities to represent concept clusters, rather than clustering based on the similarity between geometric distances of data objects like K-means clustering. Therefore, conceptual clustering has two crucial tasks: concept classification and concept generation. Beyond some standard clustering methods, Mi [20] proposed a fuzzy conceptual clustering and generation model which pays attention to both attribute information and object information. It is of great significance to improve the performance of cluster analysis and concept

* Corresponding author.
E-mail addresses: chxuwh@gmail.com (W.H. Xu), chenyaoqi17@126.com (Y. Chen).

https://doi.org/10.1016/j.knosys.2022.109472
0950-7051/

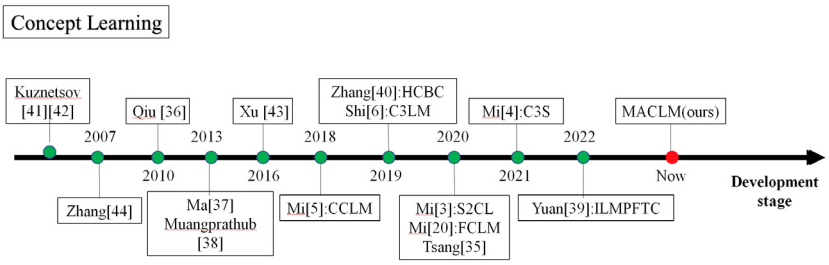

Fig. 1. Development stages of concept learning in recent years [21–30].

classification. But it needs to be pointed out that in cognitive science, humans will selectively focus on overall situation while ignoring other unconspicuous details due to the bottleneck of information processing. Humans need to select a specific part and then focus on it to rationalize the limited information-processing resources. For example, only a few words to be read will be paid attention to and processed when people are reading. However, in concept-cognitive learning, some concept learning systems proposed by scholars do not take into account attention. In addition, what we really want is for the concept space to perform well on the new samples.To achieve this goal, CCL should have the ability to identify new samples. However, when the amount of data is large, the number of concepts in the concept space will be enormous, which will bring difficulties to the identification of new samples. Therefore, clustering the concept space to generate representative concepts has important practical significance. Attention is substantial for learning a clear concept with excellent generalization ability. Motivated by these observations, we present a novel model named MA-CLM. We introduced an attention-based architecture with the aim of performing conceptual clustering in the concept space. The idea is to calculate the attention coefficient of each concept in the concept space by paying attention to its neighbors. Note that the concept clustering based on attention-architecture has several interesting characteristics: By specifying arbitrary threshold, it can be applied to conceptual clustering with different degrees. Furthermore, it can also be directly used for concept induction learning and concept generation to reduce the size of concept space and speed up concept prediction.

The main contributions of this article are listed as follows.

(1) We find that the vital role in CCL of concept selection and concept generation is attention. Therefore, attention is introduced into the CCL model. More specifically, the conceptual attention space is defined, and the concept clustering is realized by referring to the idea of graph attention. It improves the efficiency and classification accuracy of the conceptual cognitive model.

(2) Considering the information provided by the category distinguishing attribute of the pseudo-concept spaces, a new discriminant index is defined in concept category recognition, which combines maximum concept similarity and global similarity.

(3) Leveraging this framework, we conduct an extensive experimental study to evaluate the effect of the MA-CLM method on concept prediction and generation. In our experiments, MA-CLM is found to perform best for the considered datasets on classification tasks, clearly outperforming the other algorithm. Moreover, this model can also generate reasonable new concepts.

The remainder of this article is organized as follows. Some related basic knowledge are reviewed in Section 2. And we addresses the proposed method, MA-CLM, in detail, including fundamental ideas, processes, and algorithms in Section 3. In Section 4, we shows experiments and makes a comparison with S2CL$^{\alpha}$ and other classical classification algorithms on some datasets. The conclusion and our future work are proposed in Section 5.

2. Preliminaries

In this section, we briefly review some basic notions related to CCL.

Definition 1 (*[12]*)**.** A triplet (U, AT, I) is known as a formal context, where U and AT are, respectively, an object set and an attribute set, and I is a binary relation between U and AT, that is, $I \subseteq U \times AT$. Here, xIa means object x has the attribute a. Furthermore, the derivation operator is defined for $X \subseteq U, B \subseteq AT$ as follows:

$$f(X) = \{a \in AT | xIa \text{ for all } x \in X\},$$

$$g(B) = \{x \in U | xIa \text{ for all } a \in B\}.$$

$f(X)$ is the maximal set of the attributes that all the objects in X have in common and $g(B)$ is the maximal set of the objects shared by all the attributes in B. A concept in the formal context (U, AT, I) is defined to be an ordered pair (X, B) if $f(X) = B$ and $g(B) = X$, where the elements X and B of the concept (X, B) are called the extent and intent, respectively.

Property 1 (*[12]*)**.** *For any $X_1, X_2 \subseteq U$ and $B_1, B_2 \subseteq AT$, the following properties hold:*

$$X_1 \subseteq X_2 \Rightarrow f(X_2) \subseteq f(X_1)$$

$$B_1 \subseteq B_2 \Rightarrow g(B_2) \subseteq g(B_1)$$

$$f(X_1 \cup X_2) \supseteq f(X_1) \cap f(X_2)$$

$$g(B) = \{x \in U | B \subseteq f(\{x\})\}$$

Let (U, AT, I) be a formal context, 2^U and 2^{AT} be the power sets of U and AT, respectively. Then, $\mathcal{L} : 2^U \to 2^{AT}$ and $\mathcal{H} : 2^U \to 2^{AT}$ are considered as two set-valued mappings, and they are abbreviated as $\mathcal{L}$ and $\mathcal{H}$, respectively. Furthermore, if for any $X_1, X_2 \subseteq U$ and $B_1, B_2 \subseteq AT$, the following properties hold:

$$X_1 \subseteq X_2 \Rightarrow \mathcal{L}(X_2) \subseteq \mathcal{L}(X_1)$$

$$B_1 \subseteq B_2 \Rightarrow \mathcal{H}(B_2) \subseteq \mathcal{H}(B_1)$$

$$\mathcal{L}(X_1 \cup X_2) \supseteq \mathcal{L}(X_1) \cap \mathcal{L}(X_2)$$

$$\mathcal{H}(B) = \{x \in U | B \subseteq \mathcal{L}(\{x\})\}$$

Then, the two set-valued mappings $\mathcal{L}$ and $\mathcal{H}$ are referred as the cognitive operators [34]. In the cases and experiments in this paper, the operators f and g in formal concept analysis are used as cognitive operators $\mathcal{L}$ and $\mathcal{H}$.

Table 1
A regular formal decision context.

U	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	d_1	d_2
1	1	1	1	0	1	1	1	1	1	1	0
2	1	1	0	0	0	0	0	0	0	1	0
3	1	0	1	0	0	0	0	0	0	1	0
4	0	0	0	1	1	0	0	1	0	1	0
5	0	0	0	1	1	0	1	0	0	1	0
6	1	0	1	0	0	1	1	0	1	0	1
7	1	1	1	0	0	0	1	0	1	0	1
8	0	1	1	0	1	1	0	1	0	0	1
9	0	1	1	1	0	0	0	0	1	0	1

Definition 2 ([5]). Let (U, C, I) and (U, D, J) be two formal contexts, where $I \subseteq U \times C, J \subseteq U \times D$. For any $d_1, d_2 \in D$, if $\mathcal{H}(d_1) \cap \mathcal{H}(d_2) = \emptyset$, then a quintuple (U, C, I, D, J) is referred to as a regular formal decision context, where (U, C, I) and (U, D, J) are called the conditional formal context and decision formal context, respectively.

Example 1. Table 1 is the shopping records of a supermarket, in which the set $\{1, 2, \ldots, 9\}$ represents nine customers, the sets $\{c_1, c_2, \ldots, c_9\}$ and $\{d_1, d_2\}$ represent eleven commodities, and the 1 indicating that the corresponding customer has purchased the corresponding commodity. From Definition 2, we know that Table 1 expresses a regular formal decision context, where $U = \{1, 2, \ldots, 9\}$, $C = \{c_1, c_2, \ldots, c_9\}$ and $D = \{d_1, d_2\}$.

Definition 3 ([34]). For any $x \in U$ and $a \in AT$, the pairs $(\mathcal{HL}(x), \mathcal{L}(x))$ and $(\mathcal{H}(a), \mathcal{LH}(a))$ are called the granular concepts (or simply concepts) under the operators $\mathcal{L}$ and $\mathcal{H}$. Moreover, we denote the concept space that is a set of all granular concepts by $\mathcal{G}_{\mathcal{LH}}$, that is

$$\mathcal{G}_{\mathcal{LH}} = \{(\mathcal{HL}(x), \mathcal{L}(x)) | x \in U\} \cup \{(\mathcal{H}(a), \mathcal{LH}(a)) | a \in AT\}.$$

From the above formula, we know that a concept space is formed from varieties of concepts where each concept can generally be identified by two aspects: (1) extent and (2) intent [34].

Example 2. Continued with Example 1, for the class d_1, its corresponding conditional concept space $\mathcal{G}_{\mathcal{LH},d_1}$ can be shown as

$$\begin{aligned}\mathcal{G}_{\mathcal{LH},d_1} =& \{(\{1\}, \{c_1, c_2, c_3, c_5, c_6, c_7, c_8, c_9\}), (\{1, 2\}, \{c_1, c_2\}),\\ & (\{1, 3\}, \{c_1, c_3\}), (\{1, 2, 3\}, \{c_1\}), (\{1, 4\}, \{c_5, c_8\}),\\ & (\{4\}, \{c_4, c_5, c_8\}), (\{1, 5\}, \{c_5, c_7\}), (\{5\}, \{c_4, c_5, c_7\}),\\ & (\{4, 5\}, \{c_4, c_5\}), (\{1, 4, 5\}, \{c_5\})\}\end{aligned}$$

Property 2 ([5]). *For any $(X_1, B) \in \mathcal{G}^C_{\mathcal{LH}}$ and $(X_2, D_1) \in \mathcal{G}^D_{\mathcal{LH}}$, where $\mathcal{G}^C_{\mathcal{LH}}$ and $\mathcal{G}^D_{\mathcal{LH}}$ are, respectively, the concept space of (U, C, I) and (U, D, J), if $X_1 \subseteq X_2$, and X_1, B, X_2, and D_1 are nonempty, then the object set X_1 is connected with the decision attribute set D_1 under the conditional attribute set B in a regular formal decision context (U, C, I, D, J).*

Intuitively, Property 2 shows that an object $x(x \in X_1)$ can be labeled by a single label $d(d \in D_1)$.

For $(X_1, B_1), (X_2, B_2) \in \mathcal{G}_{\mathcal{LH}}$, we define the order relation $(X_1, B_1) \preccurlyeq (X_2, B_2)$ if and only if $X_1 \subseteq X_2$ (or $B_2 \subseteq B_1$). Then, we say that (X_1, B_1) is a subconcept of (X_2, B_2), and (X_2, B_2) is a super concept of (X_1, B_1).

For more details about conceptual cognitive learning, please refer to the literature [5].

3. Multi-attention concept cognitive learning

Attention is a core property of all perceptual and cognitive operations. It has now permeated most aspects of perception and cognition research. Growing consensus indicates that selection mechanisms operate throughout the brain and are involved in every stage from sensory processing to decision-making and consciousness. Attention has become a catch-all term for how the brain controls its information processing, and its effects can be measured through conscious introspection, overt and implicit behaviors, electrophysiology, and brain imaging [35]. Attentional mechanisms evolved out of necessity to efficiently focus limited processing capacity on the most important information relevant to ongoing goals and behaviors. Attention is necessary in concept-cognitive learning because concept space presents far more information than can be effectively processed at some given moment. This constraint of limited capacity is a critical factor in introducing attention.

3.1. Conceptual attention space

Exogenous, bottom-up certain special purpose draws attention to a location by a cue, such as decision-making attribute. Accordingly, in different concept spaces, the attention to the attribute is different. In the cognitive process, the attributes that are more similar to the decision attributes will be paid more attention, and the inner product is one of the methods to calculate the similarity of two vectors. Therefore, the inner product is used to calculate the attention in Eq. (1). We compute the dot products of all condition attribute vector $\{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{n_1}\}$ with decision attribute vector $\vec{d}_k (k = 1, 2, \ldots, n_2)$. In addition, for each class d_i, apply a softmax function to obtain the attentions on the attributes to make each attention be in the interval $(0, 1)$, and the attention of all attributes add up to 1. Specifically, the standard softmax function [36] $softmax : \mathbb{R}^K \to \mathbb{R}^K$ is defined by the formula

$$softmax(\vec{z})_i = \frac{exp(z_i)}{\sum_{j=1}^{K} exp(z_j)} \text{ for } i = 1, \ldots, K \text{ and } \vec{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

Example 3. Continued with Example 1, for the class d_1, the degree of attention to attribute c_1 can be calculated as

$$\begin{aligned} & softmax((\vec{c}_1 \cdot \vec{d}_1, \vec{c}_2 \cdot \vec{d}_1, \ldots, \vec{c}_9 \cdot \vec{d}_1))_1 \\ &= \frac{exp(\vec{c}_1 \cdot \vec{d}_1)}{\sum_{j=1}^{9} exp(\vec{c}_j \cdot \vec{d}_j)} \\ &= \frac{e^3}{e^3 + e^2 + e^2 + e^2 + e^3 + e^1 + e^2 + e^2 + e^1} \\ &= 0.2433. \end{aligned}$$

In practice, we compute the attention on a set of decision attributes simultaneously, packed together into a matrix $D = (\vec{d}_1, \ldots, \vec{d}_{n_2})$. The conditional attribute are also packed together into matrices $C = (\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{n_1})$. We compute the attention matrix of outputs as:

$$A = Attention(C|D) = softmax(D^T C). \quad (1)$$

where A_{ij} represents the attention of attribute c_j in the d_i class.

Example 4. Continued with Example 1, the attention matrix can be calculated as

$$A = softmax(\begin{bmatrix}1&0\\1&0\\1&0\\1&0\\1&0\\0&1\\0&1\\0&1\\0&1\end{bmatrix}^T \begin{bmatrix}1&1&1&0&1&1&1&1&1\\1&1&0&0&0&0&0&0&0\\1&0&1&0&0&0&0&0&0\\0&0&0&1&1&0&0&1&0\\0&0&0&1&1&0&1&0&0\\1&0&1&0&0&1&1&0&1\\1&1&1&0&0&0&1&0&1\\0&1&1&0&1&1&0&1&0\\0&1&1&1&0&0&0&0&1\end{bmatrix})$$

$$= softmax(\begin{bmatrix}3&2&2&2&3&1&2&2&1\\2&3&4&1&1&2&2&1&3\end{bmatrix})$$

$$= \begin{bmatrix}0.2433&0.0895&0.0895&0.0895&0.2433&0.0329&0.0895&0.0895&0.0329\\0.0591&0.1606&0.4365&0.0217&0.0217&0.0591&0.0591&0.0217&0.1606\end{bmatrix}.$$

In cognitive science, different levels of attention to attributes lead to additional attention to some concepts in the concept space. In other words, generally speaking, we pay more attention to a concept because of some salient attributes just as we are more likely to notice red objects than black ones. When we aim at a conceptual space, we inevitably focus on general issues at the expense of detail, pointing to a few relevant and unique concepts as much as possible. Inspired by this, we introduce conceptual attention space into performance concept-cognitive learning. Since a concept is composed of two parts, extent and intent, the attention to the concept depends to a certain extent on its intent. Based on this, the attention of (X, B) can be regarded as the sum of the attention of the attributes included in the conceptual intent.

Definition 4. For the class d_k, let (X, B) be a granular concept and $Attention(C|d_k) = A_{k\cdot}$ is the n_1 dimensional vector of attentions on the attributes, then the attention of (X, B) is defined as follow: $Attention((X, B)) = \vec{b}Attention(C|d_k)^T$, where $Attention(C|d_k) = (A_{k1}, A_{k2}, \ldots, A_{kn_1})$, $\vec{b} = (b_1, b_2, \ldots, b_{n_1})$, $b_i = \begin{cases}1, & c_i \in B,\\ 0, & c_i \notin B.\end{cases}$

Example 5. Continued with Example 4, for the class d_1, $(X, B) = (\{5\}, \{c_4, c_5, c_7\}) \in \mathcal{G}_{\mathcal{LH},d_1}$, the operation of attention as depicted in Fig. 2:

Furthermore, we denote conceptual attention space(CAS) which is a set of binary pairs composed of the granular concept whose attention degree exceeds a given threshold $cast(cast \in [0, 1))$ and its attention degree, that is

$$\begin{aligned}\mathcal{CAS}^{cast} = &\{((\mathcal{HL}(x), \mathcal{L}(x)), \ Attention(\mathcal{HL}(x), \mathcal{L}(x)))| x \in U,\\ &Attention(\mathcal{HL}(x), \mathcal{L}(x)) \geq cast\}\\ \cup &\{((\mathcal{H}(a), \mathcal{LH}(a)), Attention(\mathcal{H}(a), \mathcal{LH}(a)))| a \in\\ &AT, Attention(\mathcal{H}(a), \mathcal{LH}(a)) \geq cast\}.\end{aligned}$$

Example 6. Continued with Example 2, let $cast = 0.2$, for the class d_1, its corresponding conditional conceptual attention space $\mathcal{CAS}_{d_1}^{cast}$ can be shown as

$$\begin{aligned}\mathcal{CAS}_{d_1}^{cast} = &\{((\{1\}, \{c_1, c_2, c_3, c_5, c_6, c_7, c_8, c_9\}), 0.91),\\ &((\{1, 2\}, \{c_1, c_2\}), 0.33), ((\{1, 3\}, \{c_1, c_3\}), 0.33),\\ &((\{1, 2, 3\}, \{c_1\}), 0.24),\\ &((\{1, 4\}, \{c_5, c_8\}), 0.33), ((\{4\}, \{c_4, c_5, c_8\}), 0.42),\\ &((\{1, 5\}, \{c_5, c_7\}), 0.33), ((\{5\}, \{c_4, c_5, c_7\}), 0.42),\\ &((\{4, 5\}, \{c_4, c_5\}), 0.33), ((\{1, 4, 5\}, \{c_5\}), 0.24)\}.\end{aligned}$$

The hyper-parameter *cast* is the CAS reduction factor and its effect is discussed in Section 4.2.

Based on the above discussion, the complete algorithm of constructing CAS (called CCAS) is presented in Algorithm 1.

Algorithm 1: Constructing CAS

Input: A formal context (U, AT, I) and a conceptual attention space threshold *cast*.
Output: The conceptual attention space $\mathcal{CAS}^{cast}$.

```
for each x ∈ U do
    Construct a granular concept (HL(x), L(x)).
    Computing the attention of (HL(x), L(x)) by Definition 4.
    if Attention(HL(x), L(x)) ≥ cast then
        CAS^cast ⟵ ((HL(x), L(x)), Attention(HL(x), L(x))).
    end
end
for each a ∈ AT do
    Construct a granular concept (H(a), LH(a)).
    Computing the attention of (H(a), LH(a)) by Definition 4.
    if Attention(H(a), LH(a)) ≥ cast then
        CAS^cast ⟵ ((H(a), LH(a)), Attention(H(a), LH(a))).
    end
end
return: CAS^cast;
```

3.2. Conceptual clustering based on graph attention

Rosch [37] established the prototype category theory with typical samples as cognitive reference points. According to the prototype theory, a category is a concept composed of some features that are usually gathered together. These attributes are not necessary and sufficient conditions to define the category. From a cognitive perspective, all categories intersect on the edge of each other with no clear boundaries. The ambiguity and openness of the category are actually in line with the principle of cognitive economy, allowing us to use less cognitive effort to obtain as much information as possible. The original intent of the formal concept is established under sufficient and necessary conditions, which is inconsistent with the idea of the prototype theory. Based on this, we introduce the idea of conceptual clustering to further study the formal concept. When people establish or understand a category, they often take a crucial prototype as a benchmark or cognitive reference point. Ungerer and Schmid [38] call the abstract prototype not a specific sample but a general schematic representation based on category members. It is the most typical and characteristic member in the category. Archetype is a typical member of the category, and it enjoys more attributes than other members. Accordingly, we define the core concepts in the conceptual attention space as:

Definition 5. Given a subset $\mathcal{D} \subseteq \mathcal{CAS}^{cast}$, we define core concept $((X, B), Attention(X, B))$ of $\mathcal{D}$ as follows:

$$\forall((X_i, B_i), Attention(X_i, B_i)) \in \mathcal{D}, \ Attention(X_i, B_i) \leq Attention(X, B).$$

The core concept is accordance with characteristics of the archetype. Based on core concepts, we call certain concepts that have family similarities with core concepts as adjacent concepts. Here we define the concept of adjacency as a concept with the same object as the core concept, which is highly consistent with the thinking of modern category theory.

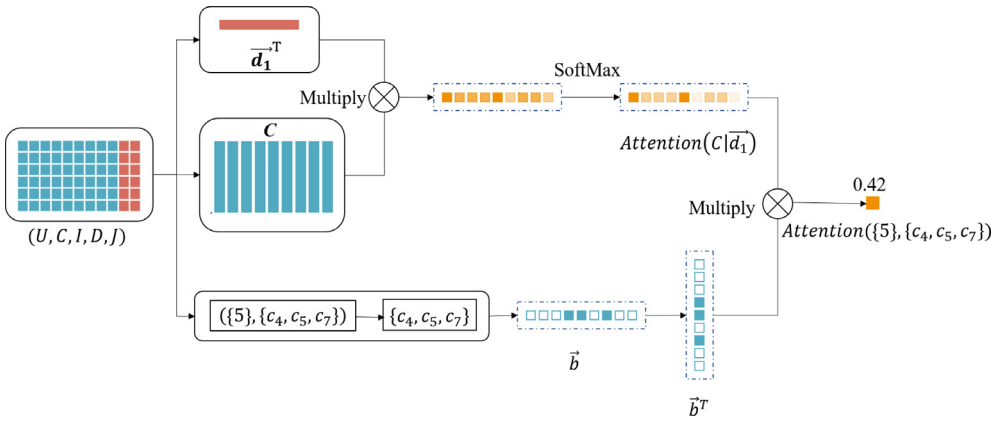

Fig. 2. An illustration of the attention operator. It first computes similarity scores between condition attribute vectors and decision attribute vector. The resulting coefficient vector is normalized by a softmax function. The attention of (X, B) is computed by taking the weighted summation over intent vectors.

Definition 6. For $((X, B), Attention(X, B)) \in \mathcal{CAS}^{cast}$, if $((X_1, B_1), Attention(X_1, B_1)) \in \mathcal{CAS}^{cast}$, and $X_1 \cap X \neq \emptyset$, then $((X_1, B_1), Attention(X_1, B_1))$ is referred to as a adjacency concept of $((X, B), Attention(X, B))$.

The value of conceptual clustering will lie in its proposal being conducive to reducing the size of concept space, making concept-cognitive learning more effective, and as a starting point for classification tasks and new concept generation. At a minimum, concept clustering serves as a portal for fast and effective classification, and at its best, it can stimulate new practical concepts and more integrative information. We introduce the concept of the concept cluster, where a pseudo-concept is used to describe a cluster. The clustering begins with core concepts, followed by the adjacency concept. A core concept and adjacency concepts characterize the structure of a concept cluster. We wish to use this structure to facilitate our concept cognition learning. Concept clustering can also be seen as the process of recognizing the concepts that have been mastered. The two most important tasks in a conceptual clustering system are concept classification and concept discovery [20]. We usually pay more attention to certain concepts when identifying in a specific concept space. Inspired by graph attention, in our clustering mechanism, pay attention to the concept with the most prominent attention value in the CAS, cluster it with its neighboring concepts together to generate pseudo-concepts, and continue this process until clustering is impossible. In order to transform the concept into more generalized pseudo-concepts with sufficient expressive power, at least one conceptual clustering transformation is required. The relevant definitions are as follows:

Definition 7. For the class d_k, let $((X, B), Attention(X, B))$ be a granular concept, $((X_1, B_1), Attention(X_1, B_1))$ be its adjacency concept and $Attention(C|d_k)$ is the vector of attentions on the attributes, then the intent attention coefficient is defined as follows:

$$e^{IN}(B, B_1) = \frac{\vec{u}Attention(C|d_k)^T}{\vec{v}Attention(C|d_k)^T},$$

where $\vec{u} = (u_1, u_2, \ldots, u_{n_1})$, $u_i = \begin{cases} 1, & c_i \in B \cap B_1, \\ 0, & c_i \notin B \cap B_1. \end{cases}$ $\vec{v} = (v_1, v_2, \ldots, v_{n_1})$, $v_i = \begin{cases} 1, & c_i \in B \cup B_1, \\ 0, & c_i \notin B \cup B_1. \end{cases}$

In addition, some standard conceptual clustering methods mainly focus on the attribute information, ignoring the object information that is also important to improve clustering analysis and concept classification ability [20]. To address the problem, we propose the extent attention coefficient.

Definition 8. Let $((X, B), Attention(X, B))$ be a granular concept, $((X_1, B_1), Attention(X_1, B_1))$ be its adjacency concept, then the extent attention coefficient is defined as follows:

$$e^{EX}(X, X_1) = \frac{|X \cap X_1|}{|X \cup X_1|}.$$

We then perform attention on the adjacency concept, computes attention coefficients

$$e((X, B), (X_1, B_1)) = iaw \cdot e^{IN} + eaw \cdot e^{EX} + (1 - iaw - eaw) \cdot Attention((X_1, B_1)), \quad (2)$$

where $iaw (iaw \in [0, 1])$, $eaw (eaw \in [0, 1])$ are the weights of the intent attention coefficient and the extend attention coefficient in the final attention coefficient, respectively. In its most general formulation, the model allows the adjacency concept with an attention coefficient greater than $cst (cst \in (0, 1))$ to attend on (X, B), dropping all structural information. We inject the graph structure into the pseudo-concept by performing attention and only compute attention e for adjacency concepts.

Let $AC^{cst} = \{((X_1, B_1), Attention(X_1, B_1)), ((X_2, B_2), Attention(X_2, B_2)), \ldots, ((X_k, B_k), Attention(X_k, B_k))\}$ be all adjacency concept with attention coefficient greater than cst of $((X, B), Attention(X, B))$. Then to make the attention coefficients easily comparable across different attention clusters, we normalize the adjacency concept attention using the softmax function. The graph attention of $((X_j, B_j), Attention(X_j, B_j))$ can be shown as

$$softmax_j(e_j) = \frac{exp(e_j)}{\sum_{j=1}^{k} exp(e_j)}, \quad (3)$$

where $e_j = e((X, B), (X_j, B_j)), ((X_j, B_j), Attention(X_j, B_j)) \in AC^{cst}$.

According to the above definitions of core concepts and adjacency concepts, we express the concept cluster as,

Definition 9. For a threshold cst, given a concept subset $\mathcal{C}^{cst} \subseteq \mathcal{D} \subseteq \mathcal{CAS}^{cast}$,we define an attention concept cluster $\mathcal{C}^{cst}$ as follow:

$$\mathcal{C}^{cst} = \{((X, B), Attention(X, B))\} \cup AC^{cst},$$

where (X, B) is a core concept of $\mathcal{D}$, AC^{cst} be the adjacency concept with attention greater than or equal to cst of (X, B) in concept set $\mathcal{D} \subseteq \mathcal{CAS}^{cast}$.

Example 7. Continued with Example 6, let $cst = 0.5$, for the class d_1, its corresponding conceptual clustering based on graph attention in $\mathcal{CAS}_{d_1}^{cast}$ can be shown as Fig. 3. The conceptual attention space can be viewed as a weighted directed graph, where the weight is the attention between concepts. Note that the attention

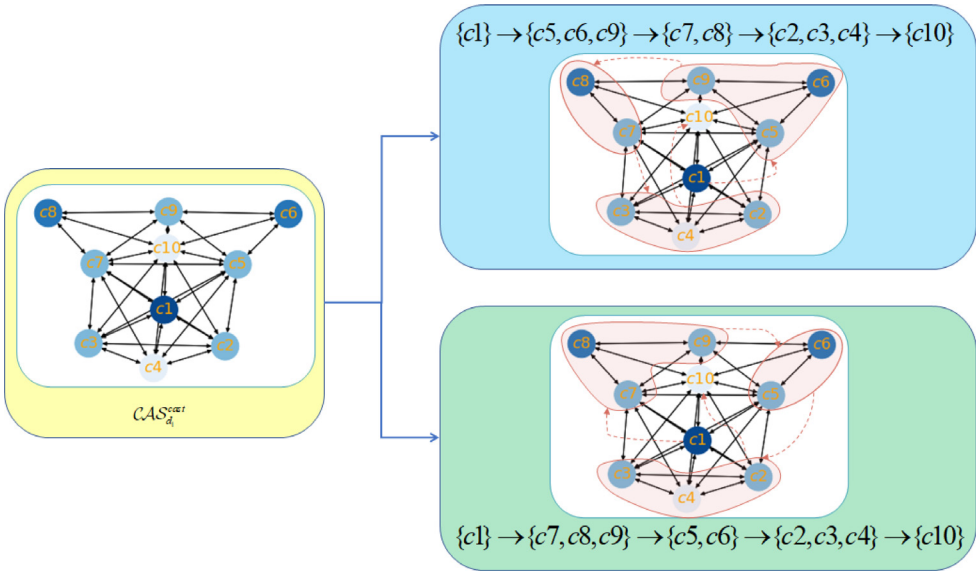

Fig. 3. An illustration of proposed conceptual clustering process based on graph attention. In this example, a conceptual attention space with ten concepts is clustered. Since there are two concepts with the same degree of attention, $c6$ and $c8$, the clustering order is different, which will produce different clustering results.

between concepts is directional. From formula (2), it can be found that $e((X, B), (X_1, B_1))$ is different from $e((X_1, B_1), (X, B))$. For the sake of brevity, we denote the conceptual attention space in Example 6 by $\mathcal{CAS}_{d_1}^{cast} = \{c1, c2, c3, c4, c5, c6, c7, c8, c9, c10\}$.

One of the benefits of attention is that they allow for dealing with concept space with redundant information, focusing on the most relevant parts of concepts. Inspired by this, we present an attention-based conceptual clustering to perform the synthesis of new concepts in concept space. The idea is to compute the hidden representations of the core concepts in the graph formed by concept space, by attending over its neighbors, following an attention strategy. The attention architecture has several interesting properties: (1) the operation is efficient to classification tasks since it compresses the number of concepts in the concept space; (2) the model is directly applicable to concept inductive learning problems, including tasks of the concept-cognition learning model generalization. Next, the definition of pseudo-concept generation based on attention is given.

Definition 10. For $\mathcal{C}^{cst}$, let $XP^{pcat} = X \cup_{i=1}^{k} X_k$ and $BP^{pcat} = B \cup \{c_i \in C | b_i^s \geq pcat\}$, where $pcat (pcat \in (0, 1])$ is the threshold of graph attention weighted summation of an attribute in the adjacency concepts of concept cluster, and

$$\vec{b}^s = (b_1^s, b_2^s, \ldots, b_{n_1}^s) = \sum_{j=1}^{k} softmax_j(e_j)\vec{b}_j,$$

where $\vec{b}_j = (b_{j1}, b_{j2}, \ldots, b_{jn_1})$, $b_{ji} = \begin{cases} 1, & c_i \in B_j, \\ 0, & c_i \notin B_j. \end{cases}$

In the adjacency concepts of concept clusters, if the weighted sum of the attention of an attribute is greater than the threshold $pcat$, this attribute is more representative of the concept cluster and belongs to the pseudo-concept intent. Then, we define that the pair (XP^{pcat}, BP^{pcat}) is a pseudo-concept induced by the $\mathcal{C}^{cst}$. In what follows, the pseudo-concept (XP^{pcat}, BP^{pcat}) is called the representation of the $\mathcal{C}^{cst}$. Pseudo-concepts can also be understood as a pair of intent and extent. Statistically speaking, the generated pseudo-concepts can characterize a new concept, but these attributes of intent are not necessary and sufficient conditions to define the extent. The attributes in the intent are shared by some objects in the extent, and the objects in the extent share some attributes in the extent, and the degree of the intent representation extent is affected by the threshold $pcat$. Note that the process of generating a new pseudo-concept is known as pseudo-concept generation.

Example 8. Continuing with Example 7, in the clustering process, after the third concept cluster is generated, the remaining concept set is $\mathcal{D} = \{c2, c3, c4, c10\}$. According to Definition 5, $c2$ is the core concept. The basic mechanism of the pseudo-concept generation process is shown in Fig. 4.

Based on Definition 10, the procedure of pseudo-concept generation is summarized in Algorithm 2.

Algorithm 2: Pseudo-concept generation

Input: A concept subset $\mathcal{D} \in \mathcal{CAS}^{cast}$, the core concept $((X, B), Attention(X, B))$, the attention threshold cst, pseudo-concept generation threshold $pcat$.
Output: The attention concept cluster $\mathcal{C}^{cst}$, the pseudo-concept (XP^{pcat}, BP^{pcat}).

$AC^{cst} = \phi$, $C^{cst} = \phi$
for $((X_i, B_i), Attention(X_i, B_i)) \in \mathcal{D}$ **do**
  if $X_i \cap X \neq \emptyset$ **then**
    Computes attention coefficients e_i of (X, B) to (X_i, B_i) by Definition 7, Definition 8,and Formula (2).
    if $e((X, B), (X_i, B_i)) > cst$ **then**
      $AC^{cst} \longleftarrow ((X_i, B_i), Attention(X_i, B_i))$
    end
  end
end
Get $softmax_i(e_i)$ by Formula (3).
$C^{cst} = \{((X, B), Attention(X, B))\} \cup AC^{cst}$
Construct a pseudo-concept (XP^{pcat}, BP^{pcat}) by Definition 9.
return: C^{cst}, (XP^{pcat}, BP^{pcat}).

Based on the above theory, the procedure of conceptual clustering based on graph attention in CAS is summarized in Algorithm 3.

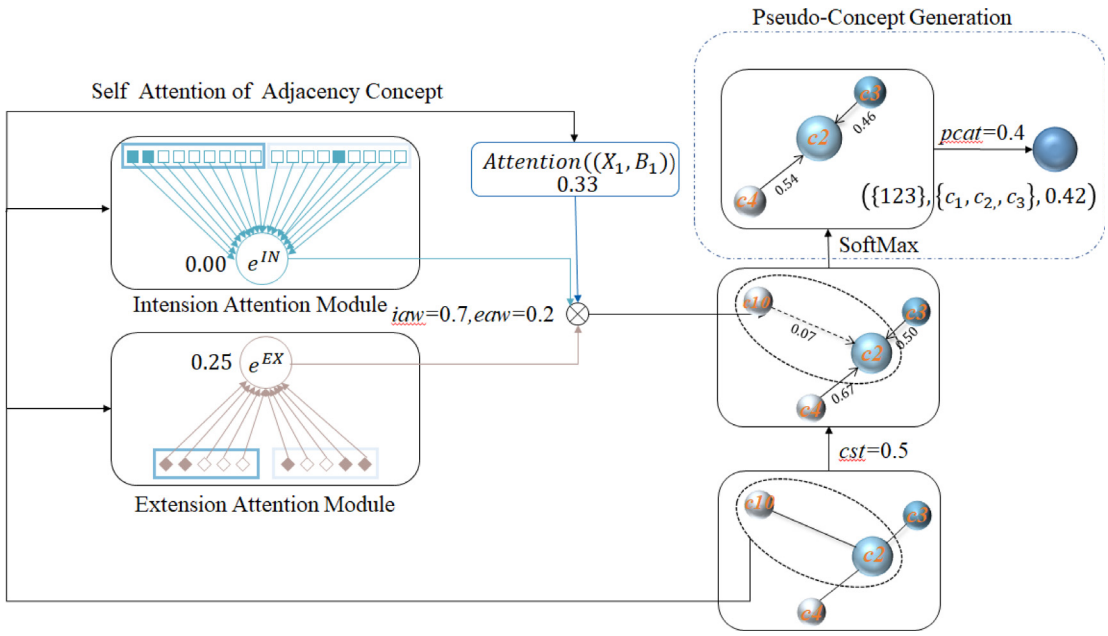

Fig. 4. An illustration of the proposed concept generation process based on graph attention. In this example, a core concept with two adjacency concepts are clustered, resulting in a pseudo-concept.

Algorithm 3: Conceptual clustering based on graph attention

Input: A conceptual attention space $\mathcal{CAS}^{cast}$, the attention threshold *cst*, pseudo-concept generation threshold *pcat*.

Output: The pseudo-concept space $\mathcal{PC}$.

$\mathcal{D} = \mathcal{CAS}^{cast}$
while $|\mathcal{D}| > 0$ **do**
  $\mathcal{C}^{cst} = \phi$
  Find the core concept $((X, B), Attention(X, B))$ of $\mathcal{D}$.
  Compute the attention concept cluster $\mathcal{C}^{c}st$ of $((X, B), Attention(X, B))$ in the concept subset $\mathcal{D}$ and the pseudo-concept (XP^{pcat}, BP^{pcat}) by Algorithm 2 .
  $\mathcal{D} = \mathcal{CAS}^{cast} \setminus \mathcal{C}^{cst}$
  $\mathcal{PC} \longleftarrow (XP^{pcat}, BP^{pcat})$
end
return: $\mathcal{PC}$;

3.3. Concept prediction

After conceptual clustering is completed in CAS, we can obtain a pseudo-concept space. Generally, different concept spaces pay different degrees of attention to attributes, which can be reflected by the distribution of different attributes in pseudo-concept spaces of different categories. Suppose an attribute frequently appears in the pseudo-concept space and rarely in other concept spaces. In that case, it is considered a unique attribute of the concept space with good classification ability and suitable for the classification task. Pseudo-concept space attribute attention is calculated by:

$$PA_{ij} = \frac{n_{ij}}{|\mathcal{PC}_i|} log \frac{\sum_i |\mathcal{PC}_i|}{\sum_i n_{ij}},$$

where n_{ij} is the number of occurrences of the attribute c_j in the d_i concept space. $|\mathcal{PC}_i|$ represent the total number of pseudo-concepts in concept space d_i. Recalculate the attention of each attribute value according to the category, and the class attribute attention matrix can be expressed as:

$$PA = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_{n_2} \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \cdots & c_{n_1} \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n_1} \\ a_{21} & a_{22} & \cdots & a_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_11} & a_{n_12} & \cdots & a_{n_2n_1} \end{pmatrix} \end{array}.$$

In the above formula, each a_{ij} value in the matrix represents the attention value of each attribute in the same category pseudo-concept space and represents the importance of attribute in the information distribution of the whole category.

For any new input sample x_i, consider it as a concept $(x, \mathcal{L}(x))$. In order to meet the needs of predicting a large amount of unlabeled data, in fact, an appropriate measure of similarity plays an important role in the learning effect. For the definition of category similarity in concept category recognition, Mi [5] considers the cognitive weight in concept similarity. Furthermore, the influence of feature difference on similarity is introduced into concept similarity, and the maximum concept similarity and average similarity in concept space are considered [3]. However, the existing definition of category similarity of new concepts only considers the similarity between the new concept and the concept in the concept space while ignoring the category similarity information of the new concept that may be provided by the category distinguishing attribute of the concept space as a whole. In light of the above problems, a concept space global similarity based on pseudo-concept space attribute attention is defined. This paper presents a mixed category similarity that combines maximum concept similarity and global similarity of concept space. A method to define the similarity degree considering the global information of concept space and the most similar concept information is proposed.

$$\begin{aligned} &Sim((x, \mathcal{L}(x)), \mathcal{PC}_i) \\ &= (1 - ga) \max_{(XP,BP)\in\mathcal{PC}_i} \{ \frac{\vec{u}\overrightarrow{PA_i}^T}{\left(\vec{u} + 2cdw \cdot \vec{v} + 2(1 - cdw) \cdot \vec{w}\right)\overrightarrow{PA_i}^T} \} \\ &\quad + ga\overrightarrow{\mathcal{L}(x)} \cdot \overrightarrow{PA_i}^T, \end{aligned} \tag{4}$$

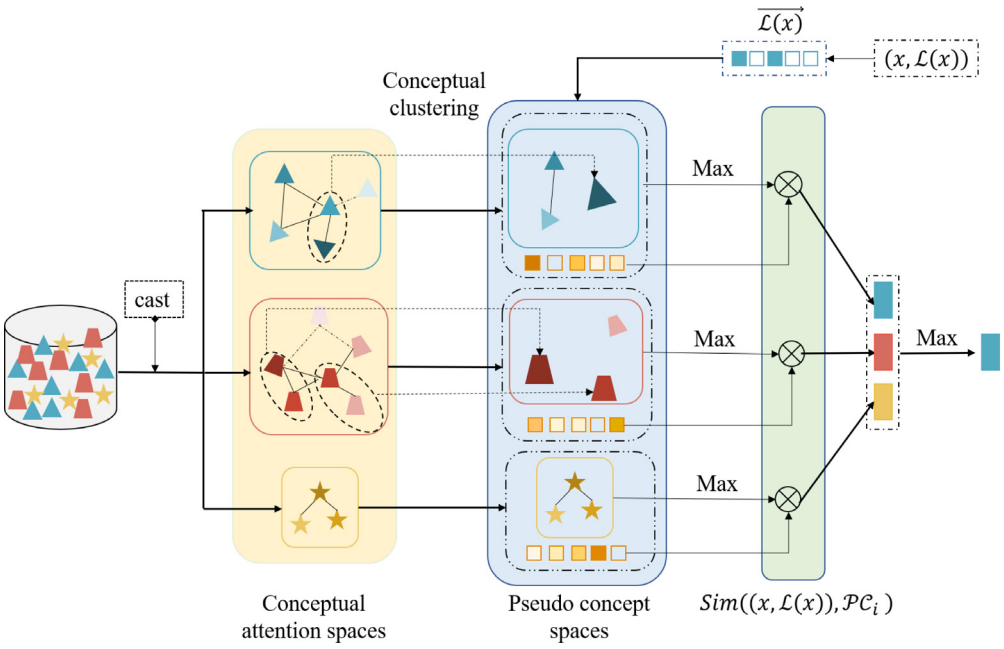

Fig. 5. Overview of our proposed MA-CLM includes three stages (Consider there are three classes to be predicted.).

where $ga (ga \in [0, 1])$ represents the weight of global similarity, the parameters $cdw (cdw \in [0, 1])$ and $(1 - cdw)$ can be, respectively, considered as the weight information added to $\mathcal{L}(x) \setminus BP$ and $BP \setminus \mathcal{L}(x)$, which express the importance of different features of $\mathcal{L}(x) \setminus BP$ and $BP \setminus \mathcal{L}(x)$ relative to the similarity degree.

$$\overrightarrow{\mathcal{L}(x)} = (b_1, b_2, \ldots, b_{n_1}), b_i = \begin{cases} 1, & c_i \in \mathcal{L}(x), \\ 0, & c_i \notin \mathcal{L}(x). \end{cases}$$

$$\vec{u} = (u_1, u_2, \ldots, u_{n_1}), u_i = \begin{cases} 1, & c_i \in \mathcal{L}(x) \cap BP, \\ 0, & c_i \notin \mathcal{L}(x) \cap BP. \end{cases}$$

$$\vec{v} = (v_1, v_2, \ldots, v_{n_1}), v_i = \begin{cases} 1, & c_i \in \mathcal{L}(x) \setminus BP, \\ 0, & c_i \notin \mathcal{L}(x) \setminus BP. \end{cases}$$

$$\vec{w} = (w_1, w_2, \ldots, w_{n_1}), w_i = \begin{cases} 1, & c_i \in BP \setminus \mathcal{L}(x), \\ 0, & c_i \notin BP \setminus \mathcal{L}(x). \end{cases}$$

3.4. Overall procedure and complexity

As shown in Fig. 5, the procedure of MA-CLM consists of three main parts: (1) constructing concept attention spaces; (2) constructing pseudo-concept space; and (3) concept generalization. In order to simplify the framework diagram without loss of generality, our framework diagram only considers a data set with three decision-making classes. In the first part, each instance type constructs a conceptual attention space based on the concept cognition operator and attribute attention. The second part is to perform conceptual clustering of the composed attention space according to Algorithm 3 to form a pseudo-concept space. In the third part, given any test instance x, obtain the binary pair $(x, \mathcal{L}(x))$ according to the concept cognition operator, and the similarity between this instance and the pseudo-concepts in each concept space will be generated according to the defined attribute attention. Then, the final prediction vector will be accomplished by aggregating the three-class vectors, and the class with maximum value will be output.

Let n, m respectively represent the number of instances and attributes in a data set. The MA-CLM is implemented on this data set. Let the time complexity of constructing a concept, computing the attention degree be $O(t_1)$, $O(t_2)$, respectively. We must recognize all objects and attributes to construct CAS, costing $O((n + m)(t_1 + t_2))$. In addition, as we showed in Algorithm 2, let the time complexity of computing the attention coefficient and constructing a pseudo-concept be $O(t_3)$, $O(t_4)$, then the time complexity of Algorithm 2 is $O(t_3 |\mathcal{CAS}| + t_4)$ in the worst case. Let the time complexity of finding the core concept be $O(t_5)$. The time complexity of running in Algorithm 3 be $O((t_3 |\mathcal{CAS}| + t_4 + t_5) |\mathcal{CAS}|)$ by Algorithm 2 in the worst case. In the third stage, we do not need to compute the similarity to all concepts in concept spaces, but only to those pseudo-concepts in pseudo-concept spaces. This means that, in practice, concept generalization requires a time of less than $O(t |\mathcal{CAS}|)$ in most cases, where $O(t)$ is the time complexity of computing the concept similarity.

3.5. Parameter tuning

We denote by $U = \{x_1, x_2, \ldots, x_n\}$ a set of instances and $K = \{1, 2, \ldots, n_2\}$ the label set. The assume that the truth instances allocation matrix Y is a $n \times k$ matrix, $Y_{ij} = 1$ means that the i_{th} instances is assigned to the j_{th} decision class. For different parameters $p = (cast, cst, pcat, eaw, iaw, cdw, ga)$, we obtain the prediction scores $\hat{Y}_{ij} = Sim((x_i, \mathcal{L}(x_i)), \mathcal{PC}_j)$ by Eq. (4). The whole model is trained by using the cross entropy classification loss as follows

$$L = -\sum_i Y_i log s(\hat{Y}_i^T),$$

where $s(\cdot)$ is the softmax function.

Because the model in this paper is difficult to get a clear analytical form of the prediction function, and numerical optimization method is not applicable, this paper uses quantum particle swarm optimization algorithm (QPSO) [39] to optimize the hyperparameters of the model.

4. Experiments

In this section, we conduct some experiments to evaluate the MA-CLM. All the experiments were executed on the computer with an Intel(R) Core(TM) i5-5200U CPU @ 2.20 GHz processor, 4 GB RAM, and a Window10 operating system. Note that our

Table 2
Characteristics of experimental datasets.

ID	Dataset	Samples	Feature	Classes	Attribute types
1	Chemical Composion of Ceramic	88	19	{44,44}	Real
2	Zoo	101	17	{41,20,5,13,4,8,10}	Categorical, Integer
3	Iris	150	4	{50,50,50}	Real
4	Wine	178	13	{59,71,48}	Integer, Real
5	Breast Cancer	286	9	{201,85}	Categorical
6	Diabetes	768	9	{500,268}	Real
7	Credit	1000	21	{300,700}	Integer, Real
8	Hypothyroid	3772	30	{194,3481,95,2}	Real
9	Mushroom	8124	22	{4208,3916}	Categorical

Table 3
Classifiers parameter tuning details.

Classifier	Parameter adjustment range	Parameter adjustment method
MA-CLM	*cast*: (0, 1], *cst*: (0, 1], *pcat*: (0, 1], *eaw*: [0, 1], *iaw*: [0, 1], *cdw*: [0, 1], *ga*: [0, 1]	Quantum particle swarm optimization
S2CL$^{\alpha}$	α: [0, 1], step size: 0.1	Grid search
DT	*max_depth*: [20, 100], step size: 10, *min_impurity_decrease*: [0.1, 0.5], step size: 0.05	Grid search
KNN	*n_neighbors*: [1, 11], step size: 1	Grid search
MNB	*alpha*: (0, 1], step size: 0.1	Grid search
SVM	*C*: [1e−3, 1e−2, 1e−1, 1, 10, 100, 1000], *gamma*: [0,1], step size: 0.1	Grid search
MLP	*hidden_layer_sizes*: [(100,), (100, 30), (50, 50), (20, 20)], *solver*: [*adam*, *sgd*, *lbfgs*], *activation*: [*identity*, *logistic*, *tanh*, *relu*]	Grid search
LR	*penalty*: [*l*1], *solver*: [*liblinear*, *saga*], *random_state*: [1, 20], step size: 1	Grid search
XGBoost	*n_estimators*: [100, 300], step size: 10, *max_depth*: [2, 15], step size: 1, *learning_rate*: [0.01, 0.1] step size: 0.01, *subsample*: [0.7, 0.9], step size: 0.02, *colsample_bytree*: [0.5, 1], step size: 0.1, *min_child_weight*: [1, 9], step size: 1	Random search
RF	*n_estimators*: [50, 300] , step size: 10, *max_depth*: [1, 20], step size: 1 *criterion*: [*gini*, *entropy*]	Random search

method and S2CL$^{\alpha}$ were implemented in a jdk8.0.1310.11 with eclipse-4.7.0 software environment and these other classification algorithms were implemented based in the Sklearn.[1]. In experiments, a total of nine data sets selected from various fields for classification in the UCI[2] have been employed for extensive comparative studies. Table 1 summarizes characteristics of each experimental data set, including the number of samples, number of features, number of samples of each class, attribute types. For the sake of brevity, we denote the nine datasets by Dataset 1–9.

The data sets in Table 2 are not formal contexts, so datasets need to be preprocessed before the experiment. For numerical genera, Kononenko's MDL Criterion [40] method is used to transform the numeric attributes in the data set into nominal attributes. Then, the nominal attributes are transformed into dummy variables. The value of the dummy variable is 0 or 1, which satisfies the definition of the formal context. At this point, the dataset can be regarded as a formal context.

4.1. Performance on test datasets in contrast with s2cl$^{\alpha}$ and other classical classification algorithms

In order to analyze the efficiency of conceptual clustering based on graph attention and global similarity between concept and concept space in the MA-CLM, S2CL$^{\alpha}$[3] with no unlabeled data, which has no conceptual clustering stage was compared. To further demonstrate the effectiveness of MA-CLM, we also performed a comparative evaluation of MA-CLM against several classical classification algorithms. The performance of MA-CLM is compared against the following eight well-established classification approaches with parameter configurations suggested in respective standard implementations. We choose the following as our baselines: Decision Tree (DT) [41] with gini index measuring the quality of the division, K-Nearest Neighbor (KNN) [41], Multinomial Naive Bayes (MNB) [41], Support Vector Machine (SVM) [41] with the Gaussian kernel function and one verse rest approach, Multi-Layer Perception (MLP) [41], Logistic Regression (LR) [42], XGBoost[43], Random Forest (RF)[44]. Meanwhile, for a fair comparison, the parameter of classifiers was tuned by a random search method or a grid search method. The details are shown in Table 3 and the default in Sklearn are used for all parameters except Table 3.

Firstly, the data set was randomly divided into the training set, validation set, and test set according to the classification ratio of 6:2:2, and the category ratio of the data set itself is maintained. On the verification set, the quantum particle swarm optimization algorithm was used to optimize the hyperparameter of MA-CLM, which was repeated 20 times, and the average value of the optimization result was taken as the final hyperparameter value. The parameter selection of the model MA-CLM in different

[1] Source codes:https://scikit-learn

[2] Available at:http://archive.ics.uci.edu/ml/datasets.html.

Table 4
The parameters of MA-CLM in datasets.

Dataset	*cast*	*cst*	*pcat*	*eaw*	*iaw*	*cdw*	*ga*
Dataset 1	0.30	0.33	0.64	0.51	0.37	0.54	0.10
Dataset 2	0.30	0.98	0.57	0.29	0.32	0.15	0.00
Dataset 3	0.28	0.98	0.29	0.84	0.74	0.53	0.00
Dataset 4	0.28	0.88	0.61	0.35	0.36	0.69	0.90
Dataset 5	0.00	0.72	0.59	0.27	0.22	0.39	0.60
Dataset 6	0.97	0.77	0.68	0.62	0.20	0.54	0.73
Dataset 7	0.30	0.61	0.61	0.52	0.45	0.56	0.59
Dataset 8	0.26	0.61	0.44	0.27	0.28	0.32	0.04
Dataset 9	0.53	0.41	0.66	0.28	0.21	0.13	0.07

Table 5
Accuracy(mean±standard deviation%) comparison with other algorithms.

Dataset	MA-CLM	S2CL[α]	DT	KNN	MNB	SVM	MLP	LR	XGBoost	RF
Dataset 1	**100.00 ± 0.00**	**100.00 ± 0.00**	99.16 ± 0.03	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**
Dataset 2	94.09 ± 2.59	94.99 ± 2.91	93.41 ± 0.19	91.36 ± 0.20	91.59 ± 0.80	86.36 ± 0.06	96.13 ± 0.08	93.64 ± 0.09	**96.36 ± 0.07**	93.86 ± 0.00
Dataset 3	**95.83 ± 3.72**	94.66 ± 3.48	95.00 ± 0.19	91.66 ± 0.32	94.83 ± 0.17	95.17 ± 0.20	94.50 ± 0.16	95.00 ± 0.19	94.33 ± 0.21	94.50 ± 0.18
Dataset 4	96.89 ± 2.52	90.67 ± 4.76	94.59 ± 0.08	99.45 ± 0.00	**99.73 ± 0.01**	99.05 ± 0.00	98.24 ± 0.05	98.65 ± 0.00	98.78 ± 0.00	98.91 ± 0.02
Dataset 5	**75.60 ± 3.80**	74.48 ± 2.13	70.69 ± 0.49	74.22 ± 0.08	73.79 ± 0.26	75.00 ± 0.05	74.19 ± 0.10	71.20 ± 0.21	70.86 ± 0.21	73.62 ± 0.13
Dataset 6	**79.31 ± 3.71**	68.86 ± 1.27	71.95 ± 0.12	76.49 ± 0.08	79.12 ± 0.12	77.53 ± 0.07	78.96 ± 0.09	78.90 ± 0.98	77.82 ± 1.25	77.01 ± 0.09
Dataset 7	**75.84 ± 1.96**	70.60 ± 2.55	70.00 ± 0.12	72.18 ± 0.04	75.03 ± 0.30	75.13 ± 0.02	73.45 ± 0.05	75.28 ± 0.04	74.80 ± 0.03	74.58 ± 0.01
Dataset 8	97.83 ± 0.37	94.49 ± 1.62	99.21 ± 0.63	97.89 ± 1.96	97.08 ± 1.93	99.11 ± 0.96	99.37 ± 0.61	**99.40 ± 0.44**	99.36 ± 0.50	99.16 ± 0.79
Dataset 9	93.72 ± 1.45	**97.41 ± 0.43**	94.59 ± 0.17	92.76 ± 0.46	89.99 ± 0.37	90.44 ± 0.46	91.77 ± 0.00	92.54 ± 0.31	92.01 ± 0.43	91.63 ± 0.82
Average_acc	**89.90**	87.35	87.62	88.45	89.01	88.64	89.62	89.40	89.37	89.25

Bold font indicates the highest accuracy.

datasets is shown in Table 4. For the same training set, the classification accuracy of the test set was calculated using the selected classifier and the proposed model. In order to avoid the contingency of experimental results, the experiment was independently repeated 20 times. Here, two widely-used metrics are utilized for performance evaluation, including mean classification accuracy and standard deviation.

The testing accuracy and standard deviation of each dataset obtained for each algorithm is reported in Tables 5. As it can be observed in Tables 5 that, the MA-CLM achieved the highest test accuracy on many datasets with relatively fine. The MA-CLM algorithm accelerates the CCL by conceptual clustering.

4.2. Parameter analysis

To illustrate the validity of parameter setting in the model and provide reference values for the future application of the model. In this section, we will analyze the influence of the parameters in the MA-CLM model on the accuracy of model classification. We analyze the model parameter groupings. Except for the parameters to be analyzed, all other parameters of MA-CLM use the values in Table 4. For each dataset, the step size of the parameters to be analyzed is set to 0.1, that is, (0.0, 0.1, . . . , 1.0). Then, for the same parameter value, 20 tests were performed independently on different training sets and test sets, and the average accuracy rate was calculated.

The parameter *cast* plays a decisive role in the number of concepts in the CAS. It can reflect the generalization degree of concepts in the conceptual space to some extent. Generally, the larger the *cast* is, the more specific the intent of concepts in the conceptual space is, and the minor generalization degree is. Different datasets have different requirements for the degree of concept generalization. Therefore, it is necessary to make analyses on the influence of changes of parameter *cast* to the accuracy of MA-CLM classification. Fig. 6 shows the trend of average accuracy of MA-CLM with parameter *cast*. We can observe from the figure that MA-CLM achieves the best accuracy for most of the datasets when the parameter *cast* is between 0.2 and 0.3. In addition, for most of datasets, it can also be observed that the accuracy of MACLM changes slowly. In this case, in order to

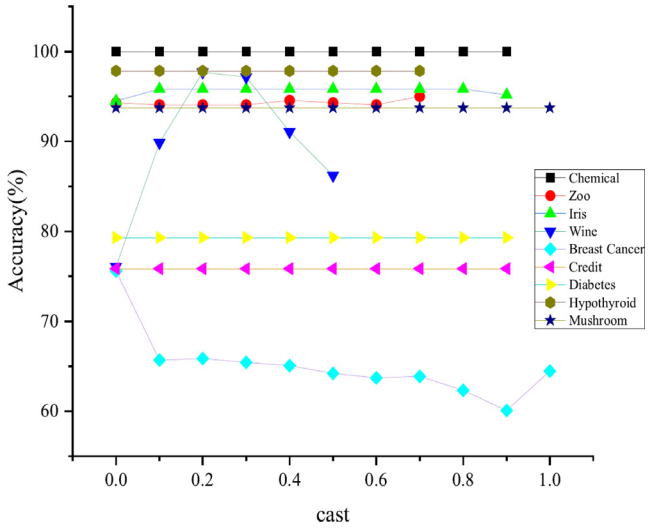

Fig. 6. The classification accuracy various with the parameter *cast*.

improve the efficiency of the model, the threshold *cast* can be increased while ensuring accuracy. Therefore, it can reasonably be inferred that it is necessary for the model to select appropriate parameters for each dataset and that the parameter *cast* in the interval [0.2, 0.3] should be worth more attention.

The parameters *cst*, *pcat* affect the clustering degree of the conceptual attention space and the pseudo-concept attribute set, respectively. Fig. 7 shows the trend of the average accuracy of MA-CLM parameters *cst*, *pact*. It can be seen from the figure that when the parameter *pcat* is between [0.1,1], MA-CLM performs well on most of datasets. Furthermore, for most of datasets, it can also be observed that the classification accuracy is lower when *pcat* is at [0,0.1] and *cst* is at [0,0.6]. Key information is lost due to the high degree of clustering, which is in line with our cognition. Therefore, the above two areas can be avoided when selecting parameters.

The parameters *iaw* and *eaw* respectively represent the weight of the intent attention and extent attention in the final attention

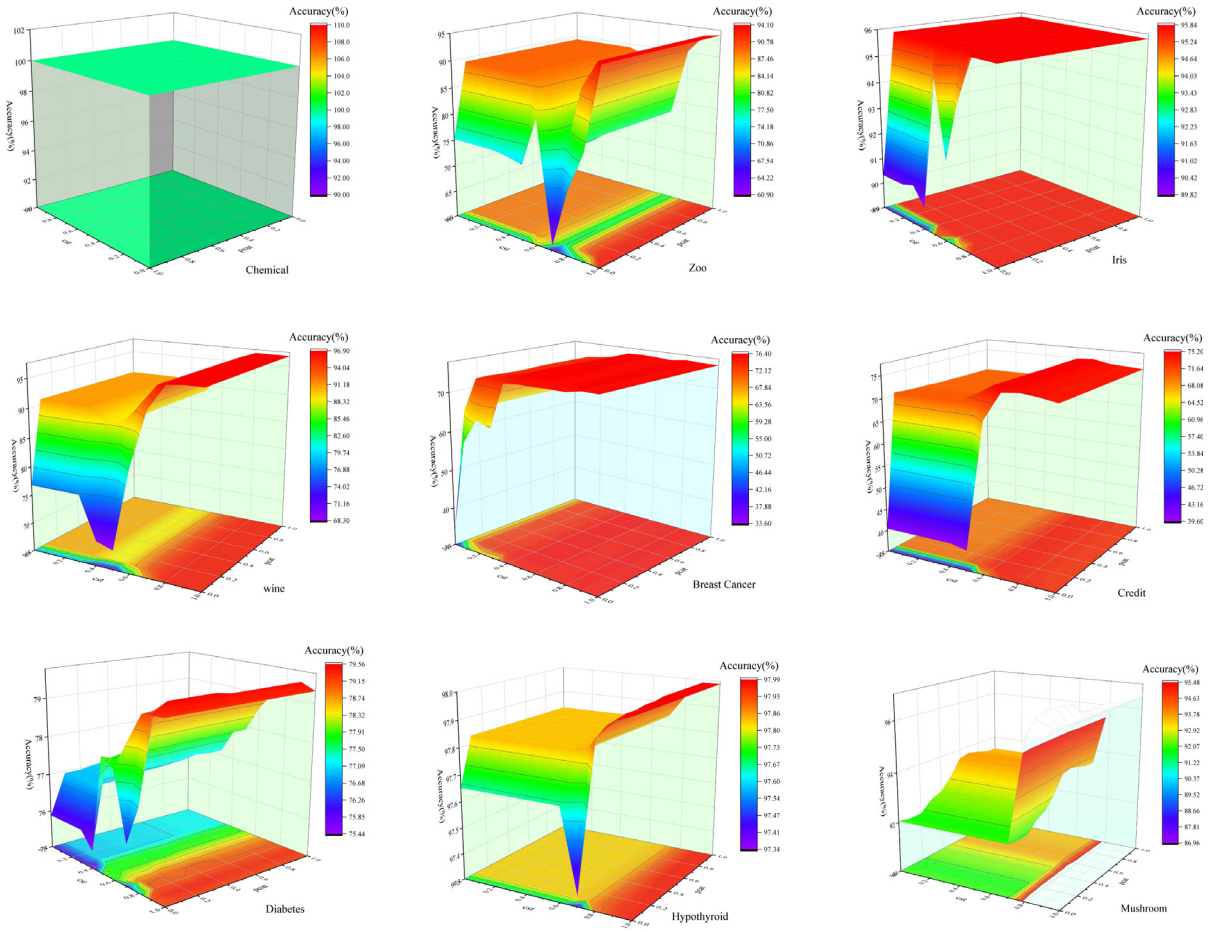

Fig. 7. The classification accuracy various with the parameter *cst* and *pcat*.

level. We pay different attention to the two in different situations. Fig. 8 shows the trend of the average accuracy of the MA-CLM parameters iaw, eaw. It can be seen from Fig. 8 that different data sets pay different attention to intent and extent. For example, the dataset Credit pays more attention to extent. When the extent attention weight is larger, the accuracy is higher. For the dataset Wine, when the intent and extent weights are small, the accuracy is lower, which shows that the intent and extent attention is significant in attention level.

The parameter cdw represents the difference between the target concept and pseudo-concept intent $\mathcal{L}(x) \setminus BP$ and $BP \setminus \mathcal{L}(x)$ to the whole impact of similarity. Fig. 9 shows the trend of the average accuracy of the MA-CLM parameter cdw. It can be seen from the figure that when the parameter cdw is between [0.2, 0.8], MA-CLM performs well on most data sets. In addition, when the parameter cdw is between [0,0.1] and [0.8,1], the accuracy of some data sets changes drastically. We can pay attention to these two thresholds when adjusting the parameters.

The parameter ga represents the influence of the category-discrimination attribute of the pseudo-concept space on the overall similarity. Fig. 10 shows the trend of the average accuracy of the MA-CLM parameter cdw. It can be seen from the figure that the accuracy of the data sets Wine, Diabetes, and Credit increases with the increase of the parameters, indicating that the category discrimination attribute provides useful information for concept prediction. In contrast, Zoo, Iris, and Hypothyroid show a downward trend. It shows that these data sets pay more attention to the maximum conceptual similarity. When the parameter cdw is in [0.2,1], most datasets have high accuracy.

4.3. Concept generation on MNIST dataset

In order to further verify the role of MA-CLM in concept generation, the MNIST[3] dataset is selected and binarized, with a threshold of 255. After that, ten samples of the number 3 are selected from the dataset for conceptual clustering and pseudo-concept generation. The model parameters are $cast = 0.07$, $cst = 0.4$, $pcat = 0.001$, $eaw = 0.5$, $iaw = 0.5$. The concept generation effect is shown in Fig. 11:

5. Conclusion

This paper mainly focuses on attention in CCL under a regular formal decision context. In other words, we have presented a multi-attention conceptual cognitive model, a novel concept-cognitive learning model that introduces intent attention, extent attention, and attention caused by attributes of the intent into the existing CCL. Moreover, considering the influence of attributes with category distinction on similarity, we give a new definition method of concept and concept space similarity.

Our model leveraging attention have successfully achieved conceptual clustering and concept generation, and a large number of experiments show that the MA-CLM model performs well in the classification task. In a word, the MA-CLM has several interesting properties. For instance, MA-CLM is efficient for classification tasks since it compresses the number of concepts in

[3] Available at:https://yann.lecun.com/exdb/mnist.

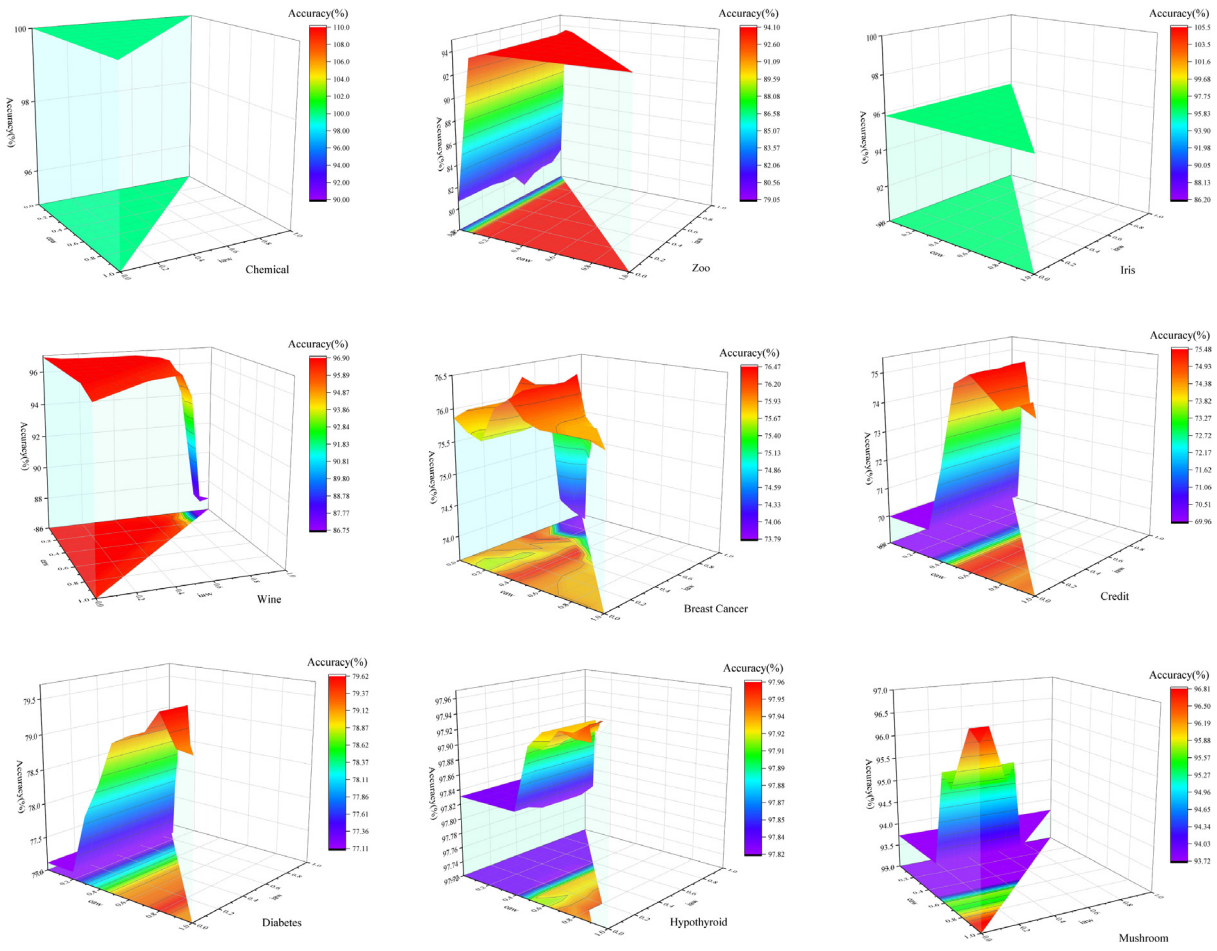

Fig. 8. The classification accuracy various with the parameter *eaw* and *iaw*.

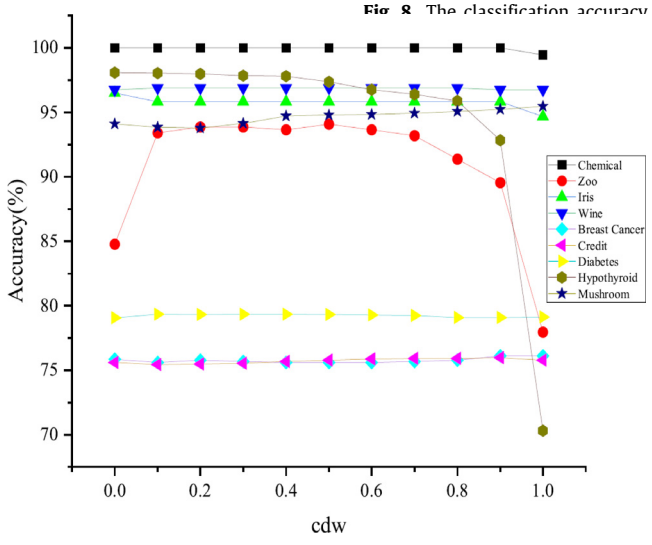

Fig. 9. The classification accuracy various with the parameter *cdw*.

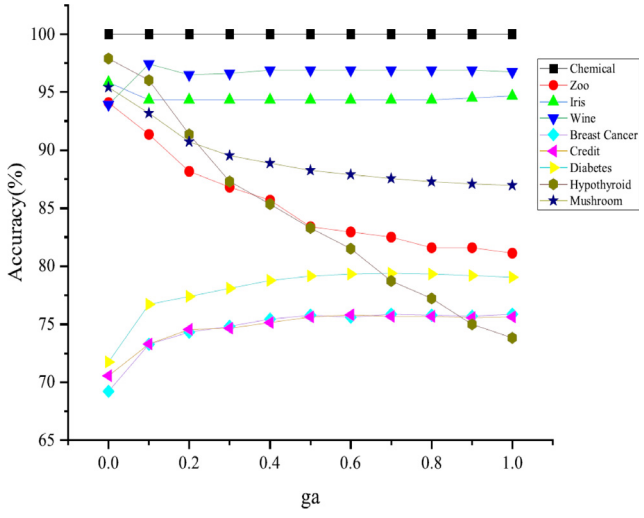

Fig. 10. The classification accuracy various with the parameter *ga*.

the concept space. It is directly applicable to concept-inductive learning, including concept generation and the task of the generalization of the concept-cognitive learning model. Besides, the information provided by the category discrimination attribute is comprehensively considered in the concept identification.

For concept-induced learning and introducing a core property in human cognitive processes, attention, into conceptual cognitive learning, we have put forward MA-CLM. However, it is still not enough in many aspects, for example, how to perform conceptual attention learning on the data stream. In addition, from our experiments, we can observe that MA-CLM cannot directly handle continuous-valued attributes. Therefore, MA-CLM for fuzzy data is also worth investigating. Our future research work will focus on these issues.

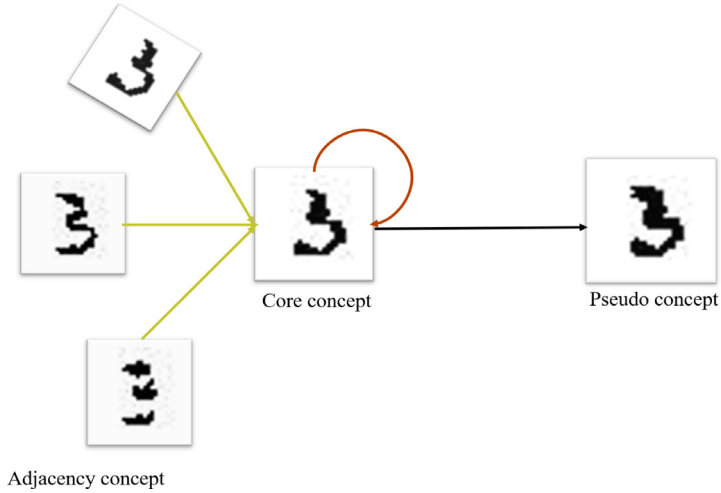

Fig. 11. The concept generation example diagram.

CRediT authorship contribution statement

Weihua Xu: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation. **Yaoqi Chen:** Data curation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 61976245) and Chongqing Postgraduate Research and Innovation Project, PR China (No. CYS21133).

References

[1] D.H. Mao, W.Z. Li, W.H. Lin, Remote sensing image classification based on formal concept analysis, J. Remote Sens. 14 (1) (2010) 90–103, http://dx.doi.org/10.3724/SP.J.1011.2010.01138.
[2] J.H. Li, Y.L. Mi, Y. Shi, Concurrent concept-cognitive learning model for classification, Inform. Sci. 496 (2019) 65–81, http://dx.doi.org/10.1016/j.ins.2019.05.009.
[3] Y.L. Mi, W.Q. Liu, Y. Shi, et al., Semi-supervised concept learning by concept-cognitive learning and concept space, IEEE Trans. Knowl. Data Eng. 34 (5) (2022) 2429–2442, http://dx.doi.org/10.1109/TKDE.2020.3010918.
[4] Y.L. Mi, P. Quan, Y. Shi, Z.R. Wang, Concept-cognitive computing system for dynamic classification, Eur. J. Oper. Res. 301 (1) (2022) 287–299, http://dx.doi.org/10.1016/j.ejor.2021.11.003.
[5] Y. Shi, Y.L. Mi, J.H. Li, W.Q. Liu, Concept-cognitive learning model for incremental concept learning, IEEE Trans. Syst. Man Cybern. 51 (2) (2021) 809–821, http://dx.doi.org/10.1109/TSMC.2018.2882090.
[6] Y. Shi, Y.L. Mi, J.H. Li, W.Q. Liu, Concurrent concept-cognitive learning model for classification, Inform. Sci. 496 (2019) 65–81, http://dx.doi.org/10.1016/j.ins.2019.05.009.
[7] G.H. Gu, Y.Y. Cao, D. Cui, et al., Object image annotation based on formal concept analysis and semantic association rules, Acta Autom. Sin. 46 (4) (2020) 767–781, http://dx.doi.org/10.16383/j.aas.c180523.
[8] M.E. Cintra, H.A. Camargo, M.C. Monard, Genetic generation of fuzzy systems with rule extraction using formal concept analysis, Inf. Sci. 349-350 (2016) 199–215, http://dx.doi.org/10.1016/j.ins.2016.02.026.
[9] M.E. Cintra, M.C. Monard, H.A. Camargo, FCA-based rule generator, a framework for the genetic generation of fuzzy classification systems using formal concept analysis, in: 2015 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, 2015, pp. 1–8, http://dx.doi.org/10.1109/FUZZ-IEEE.2015.7337950.
[10] Z. Wang, K. Hu, X. Hu, General and incremental algorithms of rule extraction based on concept lattice, Chin. J. Comput. 22 (1) (1999) 66–70.
[11] Y.Y. Yao, Interpreting concept learning in cognitive informatics and granular computing, IEEE Trans. Syst. Man Cybern. 39 (4) (2009) 855–866, http://dx.doi.org/10.1109/TSMCB.2009.2013334.
[12] R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, in: International Conference on Formal Concept Analysis, ICFCA, 2009, pp. 314–339, http://dx.doi.org/10.1007/978-3-642-01815-2_23.
[13] R. Fuentes-González, A.B. Juandeaburre, The study of the L-fuzzy concept lattice, Mathware Soft Comput. 1 (3) (1994) 209–218.
[14] Y.Y. Yao, Concept lattices in rough set theory, in: IEEE Annual Meeting of the Fuzzy Information, NAFIPS, 2004, pp. 796–801, http://dx.doi.org/10.1109/NAFIPS.2004.1337404.
[15] I. Düntsch, G. Gediga, Modal-style operators in qualitative data analysis, in: Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM, 2002, p. 155, http://dx.doi.org/10.1109/ICDM.2002.1183898.
[16] J.J. Qi, L. Wei, Y.Y. Yao, Three-way formal concept analysis, in: International Conference on Rough Sets and Knowledge Technology, RSKT, 2014, pp. 732–741, http://dx.doi.org/10.1007/978-3-319-11740-9_67.
[17] J.H. Li, C.L. Mei, Y.J. Lv, Incomplete decision contexts: Approximate concept construction rule acquisition and knowledge reduction, Internat. J. Approx. Reason. 54 (1) (2013) 149–165, http://dx.doi.org/10.1016/j.ijar.2012.07.005.
[18] L.D. Wang, X.D. Liu, Concept analysis via rough set and AFS algebra, Inform. Sci. 178 (21) (2008) 4125–4137, http://dx.doi.org/10.1016/j.ins.2008.07.004.
[19] W.Z. Wu, Y. Leung, J.S. Mi, Granular computing and knowledge reduction in formal contexts, IEEE Trans. Knowl. Data Eng. 21 (10) (2009) 1461–1474, http://dx.doi.org/10.1109/TKDE.2008.223.
[20] Y.L. Mi, Y. Shi, Y. Shi, et al., Fuzzy-based concept learning method: exploiting data with fuzzy conceptual clustering, IEEE Trans. Cybern. 52 (1) (2022) 582–593, http://dx.doi.org/10.1109/TCYB.2020.2980794.
[21] E.C.C. Tsang, B. Fan, D. Chen, et al., Multi-level cognitive concept learning method oriented to data sets with fuzziness: a perspective from features, Soft Comput. 24 (2020) 3753–3770, http://dx.doi.org/10.1007/s00500-019-04144-7.
[22] G.F. Qiu, J.M. Ma, H.Z. Yang, et al., A mathematical model for concept granular computing systems, Sci. China Inf. Sci. 53 (2010) 1397–1408, http://dx.doi.org/10.1007/s11432-010-3092-z.
[23] J.M. Ma, W.X. Zhang, Axiomatic characterizations of dual concept lattices, Internat. J. Approx. Reason. 54 (5) (2013) 690–697, http://dx.doi.org/10.1016/j.ijar.2013.01.007.
[24] J. Muangprathub, V. Boonjing, P. Pattaraintakorn, A new case-based classification using incremental concept lattice knowledge, Data Knowl. Eng. 83 (2013) 39–53, http://dx.doi.org/10.1016/j.datak.2012.10.001.
[25] K.H. Yuan, W.H. Xu, W.T. Li, W.P. Ding, An incremental learning mechanism for object classification based on progressive fuzzy three-way concept, Inform. Sci. 584 (2022) 127–147, http://dx.doi.org/10.1016/j.ins.2021.10.058.
[26] Q. Zhang, C.Y. Shi, Z.D. Niu, et al., HCBC: A hierarchical case-based classifier integrated with conceptual clustering, IEEE Trans. Knowl. Data Eng. 31 (2019) 152–165, http://dx.doi.org/10.1109/TKDE.2018.2824317.

[27] S.O. Kuznetsov, Complexity of learning in concept lattices from positive and negative examples, Discrete Appl. Math. 142 (1–3) (2004) 111–125, http://dx.doi.org/10.1016/j.dam.2003.11.002.
[28] S.O. Kuznetsov, Machine learning and formal concept analysis, in: Second International Conference on Formal Concept Analysis, ICFCA, 2004, pp. 287–312, http://dx.doi.org/10.1007/b95548.
[29] W.H. Xu, W.T. Li, Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets, IEEE Trans. Cybern. 46 (2) (2016) 366–379, http://dx.doi.org/10.1109/TCYB.2014.2361772.
[30] W.X. Zhang, W.H. Xu, Cognitive model based on granular computing, Chin. J. Eng. Math. 24 (6) (2007) 971–975.
[31] R.S. Michalski, Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data into conjunctive concepts, Int. J. Pol. Anal. Inf. Syst. 4 (1980) 219–244.
[32] Z.H. Zhou, Machine Learning, Tsinghua University. Press, Beijing, China, 2016.
[33] C. Zhang, Y.P. Kwon, J. Kramer, et al., Deep learning for design in concept clustering, in: The ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, DETC, 2017, pp. 6–9, http://dx.doi.org/10.1115/DETC2017-68352.
[34] J.H. Li, C.L. Mei, W.H. Xu, et al., Concept learning via granular computing: A cognitive viewpoint, Inform. Sci. 298 (2015) 447–467, http://dx.doi.org/10.1016/j.ins.2014.12.010.
[35] H. Pashler, J.C. Johnston, E. Ruthruff, Attention and performance, Annu. Rev. Psychol. 52 (1) (2001) 629–651, http://dx.doi.org/10.1146/annurev.psych.52.1.629.
[36] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, 2016, http://www.deeplearningbook.org.
[37] E. Rosch, C.B. Mervis, Family resemblances: Studies in the internal structure of categories, Cogn. Psychol. 7 (4) (1975) 573–605, http://dx.doi.org/10.1016/0010-0285(75)90024-9.
[38] F. Ungerer, H. Schmid, An Introduction to Cognitive Linguistics, second ed., Foreign Language Teaching and Research Press, 2008, http://dx.doi.org/10.4324/9781315835396.
[39] J. Sun, W. Xu, B. Feng, A global search strategy of quantum-behaved particle swarm optimization, in: IEEE Conference on Cybernetics and Intelligent Systems, ICCIS, 2004, pp. 111–116, http://dx.doi.org/10.1109/ICCIS.2004.1460396.
[40] K. Lgor, On biases in estimating multi-valued attributes, in: Proceedings of 14th Internationa Joint Conference on Articial Intelligence, IJCIA, 1995, pp. 1034–1040.
[41] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.
[42] D.R. Cox, Corrigenda: The regression analysis of binary sequences, J. Roy. Stat. Soc. 20 (2) (1958) 215–242.
[43] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.
[44] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.