ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

Collaborative Reduction of Features and Instances in the Set-valued Decision System

Ruimin Li[1,a,*]

[1]*School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo City, China*
[a]*liruimin@home.hpu.edu.cn*
[*]*Corresponding author*

Abstract: *Due to the advancement of technology, data is becoming richer in features and instances, but not all features and instances help to improve classification performance in data mining. Data reduction helps to alleviate the difficulty of learning techniques when the data is large, and rough sets have been widely used for data reduction. Semi-monolayer covering rough set is an efficient and high-quality rough set model in set-valued information systems. In this paper, a new data reduction scheme is proposed from the perspective of incremental updating of semi-monolayer covering rough set (abbr. FSMCDE). Firstly, in the set-valued decision system, based on the fact that the lower approximation set gradually increases with the features until it remains stable, the limit for the lower approximation set of semi-monolayer covering rough set to remain stable is proved, and the incremental updating theory of the lower approximation set is designed. Secondly, the features are continuously added to the set-valued system, and the incremental algorithm is used to update the lower approximation set until it reaches the limit, completing the collaborative reduction of features and instances. Furthermore, to reduce the blindness of adding features during incremental updating, Fisher score is introduced to form the final collaborative reduction algorithm of features and instances. The experimental results show that FSMCDE can efficiently accomplish the collaborative reduction of features and instances, and effectively improve the classification performance.*

Keywords: *Set-valued information system; Semi-monolayer covering; Incremental updating; Collaborative reduction; Fisher Score*

1. Introduction

With the rapid development of technology and the dramatic increase in data collection capabilities, data are becoming richer in dimensionality and size (number of instances). Too high dimensionality and noise in data strengthen the classification model deviation and increase the computational complexity in data mining [1]. Therefore, the demand for data reduction is increasing.

Data reduction is a data preprocessing task. The data reduction process in data mining mainly includes feature selection and instance selection[2, 3]. Feature selection can identify redundant features, and instance selection can eliminate misleading training instances. To develop a more effective model, the collaborative reduction of design features and instances can be considered.

Rough set theory[4] is a mathematical tool for data analysis and knowledge discovery. Feature selection and instance selection are two important application fields of rough set theory. Feature selection based on rough set has been carried out a lot of work, and has been successfully applied in bioinformatics and other fields[5]. Compared to the large amount of research on feature selection, there is less work on instance selection[6, 7]. These instance selection algorithms tend to use the lower approximation set to select instances that are compatible with the decision results. In studying simultaneous feature and instance selection, most algorithms based on rough set mainly focus on intelligent optimization algorithms. Anaraki et al.[8] proposed a select features and instances simultaneously method based on shuffled frog leaping algorithm. Derrac et al. [9]proposed a hybrid evolutionary algorithm for data reduction, carried out by a steady-state genetic algorithm. Although the collaborative reduction algorithm based on rough set features and instances can reduce the time cost and effectively improve the classification performance, there is not much research in this area.

In the decision system, the lower approximation set gradually expands with the increase of features. When enough features are added, the lower approximation set remains stable. This provides a new idea

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

for the collaborative reduction of features and instances. To achieve the stable state of the approximate set more quickly, incremental updating is an effective means. Incremental methods in set-valued information systems have been studied extensively[10]. Recently, Wu et al.[11, 12] proposed a semi-monolayer covering rough set for set-valued information systems. The model is superior to other rough set models based on tolerance relation in approximation quality and computational efficiency. However, in semi-monolayer covering rough set, the incremental updating theory of approximation set under feature change needs to be further studied.

In summary, based on the incremental updating theory of semi-monolayer covering rough set, this paper designs a collaborative reduction algorithm of features and instances (abbr. FSMCDE). Specifically, for semi-monolayer covering approximation operator DE0, this paper describes the increasing trend of the DE0 lower approximation set with the increase of features, proves the limit of the DE0 lower approximation set, and designs the incremental updating theory of the DE0 lower approximation set. Based on the incremental update theory and incremental limit, the features are continuously added to the set-valued system, and the incremental algorithm is used to quickly update the DE0 lower approximation set until the limit is reached, thereby completing the collaborative reduction of features and instances. Experiments show that FSMCDE can efficiently achieve the collaborative reduction of features and instances, improve data quality, and effectively improve classification performance.

2. Preliminaries

Let $S=(U,A,V,f)$ be a set-valued information system (abbr. SVIS), where U is the universe, A is a finite set of attributes, $V=\{\cup_{a\in A}V_a\}$ is the attribute value domain, and $f{:}U\times A\to 2^{V_A}$ represents a set-valued mapping from $U\times A$ to V. $DS=(U,A\cup d,V,f)$ is a set-valued decision system (abbr. SVDS), where d represents the decision attribute values and $A\cap d=\phi$. The decision set is a partition of U, $U/d=\{D_i|x\in U,|\vec{x}|_d=|\vec{y}|_d\}$.

Definition 1. [11, 12] Let U be an universe, and C be a representative covering on U. If every $K\in C$ is indispensable, i.e. $\cup\{K^{'}|K^{'}\in C,K^{'}\neq K\}\subset U,C=\{K_1,K_2,...K_n\}$ is a semi-monolayer covering on U (abbr. SMC).

- x is a reliable element of K, if $\forall L\in C(x\in L\Rightarrow K=L)$. The set of all reliable elements in U is U_0.
- K_0 is the reliable set of K, which contains all of the reliable elements in K.
- x is controversial element, if $\exists K_1,K_2\in C$, $x\in K_1$ and $x\in K_2$.

Definition 2.[13] Let $S=(U,A,V,f)$ be SVIS, where $A=\{a_1,a_2,...,a_n\}$. The information of an object $x\in U$ in S is a vector, $\vec{x}=<f(x,a_1),f(x,a_2),...,f(x,a_n)>$.

Definition 3.[12] Let $S=(U,A,V,f)$ be SVIS. $Cell_x=\{y\in U|\vec{x}=\vec{y}\}$ is a set of elements with same information explanation on S as x. The set of all cells on S is denoted by $CELL$. $CELL$ is a partition of universe. The information explanation $|Cell_x|$ of $Cell_x$ is $\overrightarrow{Cell_x}=\vec{x}$, where $x\in Cell$.

- If every value in $\overrightarrow{Cell}$ is a single-valued, the cell is called reliable cell. The reliable cell is denoted by $Cell_r$, and the set of the reliable cells is denoted by RC.
- The related reliable cell set of controversial cell $Cell_c$ is denoted by $RS(Cell_c)$, where $RS(Cell_c)=\{Cell_r\in RC|\forall a_i\in A,\ x\in Cell_m,\ y\in Cell_n,\ f(x,a_i)\subseteq f(y,a_i)\}$.
- If there exists any value in $\overrightarrow{Cell}$ is a set-valued, the cell is called controversial cell. The controversial cell is denoted by $Cell_c$, and the set of controversial cells is denoted by CC.

In semi-monolayer covering approximation space, we will introduce DE0 approximation operator in Theorem 1. DE0 not only improve the approximation quality but also accelerate the calculation of approximation set.

Theorem 1.[12] Let (U,A,V,f) be SVIS. For any $X\subseteq U$, the lower DA0 approximation set of X on SVIS is as follows, where $\underline{C}_{GC0}(X)=\{Cell_r\in RC|Cell_r\subseteq X\}$.

$$\underline{C}_{DE0}(X)=\cup\{Cell\in CELL|RS(Cell)\cap\underline{C}_{GC0}(X)\neq\varnothing\}$$

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

3. The Collaborative Reduction of Features and Instances Based on Incremental Method

Definition 4. Let $S=(U,A,V,f)$ be a set-valued information system (SVIS). $P,Q\subseteq A$ and $P\cap Q=\emptyset$. $S^P=(U,P,V,f)$ and $S^{P\cup Q}=(U,P\cup Q,V,f)$ are two subsystems of S.

● The truncation of information explanation of $Cell$ on P is denoted as $|\overrightarrow{Cell}|_P$. The information explanation $\overrightarrow{Cell^P}$ in S^P is the abbreviation of $|\overrightarrow{Cell^P}|_P$.

● $Cell^{P\cup Q}\in S^{P\cup Q}$ is the segmentation of $Cell^P\in S^P$ iff $|\overrightarrow{Cell^{P\cup Q}}|_P=\overrightarrow{Cell^P}$.

Lemma 2. Let $S^P=(U,P,V,f)$ and $S^{P\cup Q}=(U,P\cup Q,V,f)$ be the set-valued information systems. $Cell^{P\cup Q}\in S^{P\cup Q}$ is the segmentation of $Cell^P\in S^P$ iff $Cell^{P\cup Q}\subseteq Cell^P$. Furthermore, $Cell^P$ is the only one for $Cell^{P\cup Q}$ in S^P.

Proof. According to the Definition 3, $Cell_x^{P\cup Q}=\{y\in U||\vec{x}|_{P\cup Q}=|\vec{y}|_{P\cup Q}\}=\{y\in U||\vec{x}|_P=|\vec{y}|_P,|\vec{x}|_Q=|\vec{y}|_Q\}$ and $Cell_x^P=\{y\in U||\vec{x}|_P=|\vec{y}|_P\}$. "$\Rightarrow$" Obviously, $Cell_x^{P\cup Q}$ is the subset of $Cell_x^P$. "$\Leftarrow$" Any subset of $Cell^P$ has the same segmentation of information explanation. If the $Cell_x^{P\cup Q}$ is a subset of $Cell_x^P$, $Cell_x^{P\cup Q}$ is the segmentation of $Cell_x^P$ according to Definition 3. On the other hand, the truncation of information explanation of $Cell^{P\cup Q}$ is specific. Therefore, the $Cell^P$ is the only one in S^P without any question.

Theorem 3. Let $S=(U,A,V,f)$ be a set-valued information system (SVIS). $Cell\in CELL$, and $RS(Cell)=\{Cell_r||Cell_r|\preceq|Cell|,Cell_r\in RC\}$.

a) For any $a\in A$, $|\overrightarrow{Cell_r}|_a\subseteq|\overrightarrow{Cell}|_a$ and $|\overrightarrow{Cell}|_a=\cup\{|\overrightarrow{Cell_r}|_a|Cell_r\in RS(Cell)\}$.

b) If $Cell_r\in RC$ and $Cell_r\notin RS(Cell)$, there exists $a_0\in A$ satisfying that $|\overrightarrow{Cell_r}|_{a_0}\cap|\overrightarrow{Cell}|_{a_0}=\emptyset$.

Proof. According to Definition 3, $RS(Cell)=\{Cell_r\in RC||Cell_r|\preceq|Cell|\}=\{\forall a_i\in A,x\in Cell_m,y\in Cell_n,f(x,a_i)\subseteq f(y,a_i)\}$. $|\overrightarrow{Cell}|_a$ is $f(x,a)$. Thus a) is clear. On the other hand, we have noticed that $Cell_r$ is reliable. For any $a\in A$, $|\overrightarrow{Cell_r}|_a$ is a single-valued set. Therefore, $|\overrightarrow{Cell_r}|_a\subseteq|\overrightarrow{Cell}|_a\Leftrightarrow|\overrightarrow{Cell_r}|_a\cap|\overrightarrow{Cell}|_a\neq\emptyset$. If there does not exist $a_0\in A$, $Cell_r\in RS(Cell)$. It is a contradiction with $Cell_r\notin RS(Cell)$. Thus, there exists $a_0\in A$ satisfying that $|\overrightarrow{Cell_r}|_{a_0}\cap|\overrightarrow{Cell}|_{a_0}=\emptyset$, b) is also clear.

Theorem 4. Let $S^P=(U,P,V,f)$ be SVIS and $S^{P\cup Q}=(U,P\cup Q,V,f)$ be the segmentation of S^P. For $Cell^P\in S^P$ and $Cell^{P\cup Q}\in S^{P\cup Q}$, $RS(Cell^P)=\{Cell_r^P|\overrightarrow{Cell_r^P}\preceq\overrightarrow{Cell^P}\}$ and $RS(Cell^{P\cup Q})=\{Cell_r^{P\cup Q}|\overrightarrow{Cell_r^{P\cup Q}}\preceq\overrightarrow{Cell^{P\cup Q}}\}$. Suppose that $Cell^{P\cup Q}$ is the segmentation of $Cell^P$.

a) For any $Cell_r^{P\cup Q}\in RS(Cell^{P\cup Q})$, there exists $Cell_{r_0}^P\in RS(Cell^P)$ satisfying that $Cell_r^{P\cup Q}\subseteq Cell_{r_0}^P$.

b) For any $Cell_r^P\in RS(Cell^P)$, there exists $Cell_{r_0}^{P\cup Q}\in RS(Cell^{P\cup Q})$ satisfying that $Cell_{r_0}^{P\cup Q}\subseteq Cell_r^P$.

Proof. According to Lemma 2,"$Cell^{P\cup Q}$ is the segmentation of $Cell^P$ "equals to "$Cell^{P\cup Q}\subseteq Cell^P$".

a) For any $Cell_r^{P\cup Q}\in RS(Cell^{P\cup Q})$, let $|\overrightarrow{Cell_r^{P\cup Q}}|_P=\overrightarrow{Cell_{r_0}^P}$. Suppose that $Cell_{r_0}^P\notin RS(Cell^P)$. Then there exists $a_0\in P$ satisfying that $|\overrightarrow{Cell_r^{P\cup Q}}|_{a_0}\cap|\overrightarrow{Cell^{P\cup Q}}|_{a_0}=\emptyset$ (Theorem 3-b). $|\overrightarrow{Cell^{P\cup Q}}|_{a_0}\neq|\overrightarrow{Cell^P}|_{a_0}$ (Theorem 3-a). It is a contradiction with "$Cell^{P\cup Q}\subseteq Cell^P$". Therefore, $Cell_{r_0}^P\in RS(Cell^P)$ and $Cell_r^{P\cup Q}\subseteq Cell_{r_0}^P$.

b) Select any $Cell_r^P$ from $RS(Cell^{P\cup Q})$. Suppose that for any $Cell_r^{P\cup Q}\subseteq Cell_r^P$, $Cell_r^{P\cup Q}\in RS(Cell^{P\cup Q})$. According to Lemma 2, if $|\overrightarrow{Cell_r^{P\cup Q}}|_P=\overrightarrow{Cell_r^P}$, $Cell_r^{P\cup Q}\subseteq Cell_r^P$. Thus $\overrightarrow{Cell_r^P}\not\preceq|\overrightarrow{Cell^{P\cup Q}}|_P$. It is a contradiction with $|\overrightarrow{Cell^{P\cup Q}}|_P=\overrightarrow{Cell^P}\succeq\overrightarrow{Cell_r^P}$. Therefore, there exists $Cell_{r_0}^{P\cup Q}\in RS(Cell^{P\cup Q})$ satisfying that $Cell_{r_0}^{P\cup Q}\subseteq Cell_r^P$.

In $S^{P\cup Q}$, the cells are denoted as $Cell^{P\cup Q}$ and the lower DE0 approximate set of X is denoted as $\underline{C}_{DE0}^{P\cup Q}(X)$. We will discuss the relationship between $\underline{C}_{DE0}(X)$ in S^P and $S^{P\cup Q}$.

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

Theorem 5. Let $S^{P\cup Q}$ be SVIS and it also be the segmentation of S^P. Suppose $Cell^P \in S^P$ and $Cell^{P\cup Q} \in S^{P\cup Q}$. For any $X \in U$, if $Cell^P \subseteq \underline{C}^P_{DE0}(X)$, then $Cell^P \subseteq \underline{C}^{P\cup Q}_{DE0}(X)$.

Proof. Denote that the set of segmentation of $Cell^P$ in $S^{P\cup Q}$ is $A=\{Cell^{P\cup Q}|Cell^{P\cup Q}\subseteq Cell^P\}$, and $Cell^P = \cup_A(Cell^{P\cup Q})$. If $Cell^P \subseteq \underline{C}^P_{DE0}(X)$, there exists $Cell^P_{r_0} \in RS(Cell^P)$ satisfying that $Cell^P_r \subseteq X$. $Cell^{P\cup Q} \in A$ and $Cell^{P\cup Q} \subseteq Cell^P$. According to Theorem 4-b, for any $Cell^P_r \in RS(Cell^P)$, there must exist$Cell^{P\cup Q}_{r_0} \in RS(Cell^{P\cup Q})$ satisfying that $Cell^{P\cup Q}_{r_0} \subseteq Cell^P_r$. Thus for the $Cell^P_{r_0}$, there has $Cell^{P\cup Q}_{r_1}$ satisfying that $Cell^{P\cup Q}_{r_1} \subseteq Cell^P_{r_0} \subseteq X$. There has always at least one $Cell^{P\cup Q}_{r_1} \in RS(Cell^{P\cup Q})$ and $Cell^{P\cup Q}_{r_1} \subseteq X$. Therefore, $Cell^{P\cup Q} \subseteq \underline{C}^{P\cup Q}_{DE0}(X)$. Based on the arbitrariness of $Cell^{P\cup Q} \in A$, $Cell^P \subseteq \underline{C}^{P\cup Q}_{DE0}(X)$.

Corollary 6. Let $S^{P\cup Q}$ be SVIS and it also be the segmentation of S^P. For any $X \in U$, $\underline{C}^P_{DE0}(X) \subseteq \underline{C}^{P\cup Q}_{DE0}(X)$.

Proof. $\underline{C}^P_{DE0}(X) \subseteq \underline{C}^{P\cup Q}_{DE0}(X)$ are the direct conclusions of Theorem 5. $\underline{C}^{P\cup Q}_{DE0}(X)$ and the additional cells in $S^{P\cup Q}$ can be found following the conclusions in Theorem 5.

The incremental theory of $\underline{C}_{DE0}(X)$ can be found in Theorem 7.

Theorem 7. Let $S^{P\cup Q}$ be a set-valued information system and it also be the segmentation of S^P. For any $X \in U$, $\underline{C}^{P\cup Q}_{DE0}(X) = \underline{C}^P_{DE0}(X) \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$。where$\triangle LGC0^{P\cup Q}(X) = \{Cell^{P\cup Q}_r | Cell^{P\cup Q}_r \subseteq X, Cell^{P\cup Q}_r \subseteq Cell^P_r, Cell^P_r \nsubseteq X\}$.

Proof. Firstly, we need prove that $\underline{C}^{P\cup Q}_{DE0}(X) \cap \left(\underline{C}^P_{DE0}(X)\right)^C \subseteq \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$. Let $Cell^{P\cup Q}$ be any one in $\underline{C}^{P\cup Q}_{DE0}(X) \cap \left(\underline{C}^P_{DE0}(X)\right)^C$ and $Cell^P$ be the unique cell in S^P satisfying that $Cell^{P\cup Q} \subseteq Cell^P$. $Cell^{P\cup Q} \subseteq \underline{C}^P_{DE0}(X)$ means that there has $Cell^{P\cup Q}_{r_0} \in RS(Cell^{P\cup Q})$ satisfying that $Cell^{P\cup Q}_{r_0} \subseteq X$. By Theorem 4-b, there exists $Cell^P_{r_0} \in RS(Cell^P)$ and $Cell^{P\cup Q}_{r_0} \subseteq Cell^P_{r_0}$. On the other hand, $Cell^{P\cup Q} \nsubseteq Cell^P_{DE0}(X)$ means that $Cell^P \nsubseteq Cell^P_{DE0}(X)$. Therefore, every $Cell^P_r \in RS(Cell^P)$ is not the subset of X. Therefore, $Cell^P_{r_0} \nsubseteq X$ either. It means that $Cell^{P\cup Q}_{r_0} \subseteq X$, $Cell^P_{r_0} \subseteq X$ where $Cell^{P\cup Q}_{r_0} \subseteq Cell^P_{r_0}$. $Cell^{P\cup Q}_{r0} \in \Delta LGC0^{P\cup Q}(X)$.$Cell^{P\cup Q}_{r0} \in RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X)$.

Therefore, $\underline{C}^{P\cup Q}_{DE0}(X) \cap \left(\underline{C}^P_{DE0}(X)\right)^C \subseteq \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$, and $\underline{C}^{P\cup Q}_{DE0}(X) \supseteq \underline{C}^P_{DE0}(X) \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$.

Secondly, let $Cell^{P\cup Q}_r \in RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X)$. Because $Cell^{P\cup Q}_r \in \Delta LGC0^{P\cup Q}(X)$, $Cell^{P\cup Q}_r \subseteq X$. And$Cell^{P\cup Q}_r \in RS(Cell^{P\cup Q})$. It is clear that $RS(Cell^{P\cup Q}) \cap \underline{C}^{P\cup Q}_{GC0}(X) \neq \varnothing$. Therefore, $\underline{C}^{P\cup Q}_{DE0}(X) \supseteq \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$, $\underline{C}^{P\cup Q}_{DE0}(X) \supseteq \underline{C}^P_{DE0}(X) \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$(Corollary 6).

Therefore, $\underline{C}^{P\cup Q}_{DE0}(X) \supseteq \underline{C}^P_{DE0}(X) \cup \{Cell^{P\cup Q}|RS(Cell^{P\cup Q}) \cap \Delta LGC0^{P\cup Q}(X) \neq \varnothing\}$.

The incremental limit of $\underline{C}_{DE0}(X)$ can be found in Theorem 8.

Theorem 8. Let S^P be SVIS. For any $X \in U$, if $\underline{C}^P_{GC0}(X) = \overline{C}^P_{GC0}(X)$, then no matter how many segmentations, $\underline{C}_{DE0}(X)$ no longer change, where $\underline{C}_{GC0}(X) = \{Cell_r \in RC|Cell_r \subseteq X\}$ and $\overline{C}_{GC0}(X) = \{Cell_r \in RC|Cell_r \cap X \neq \varnothing\}$.

Proof. If $\underline{C}^P_{GC0}(X) = \overline{C}^P_{GC0}(X)$, then as long as $Cell^P_r$ that satisfies $Cell^P_r \cap X \neq \varnothing$, there is $Cell^P_r \subseteq \underline{C}^P_{GC0}(X)$, which means keep adding attributes and there will be no new $Cell^{P\cup Q}_{r_0} \subseteq \underline{C}^{P\cup Q}_{GC0}(X)$. that is $\underline{C}^{P\cup Q}_{GC0}(X) = \underline{C}^P_{GC0}(X)$, $\Delta LGC0^{P\cup Q}(X) = \varnothing$. According to the theorem 7, $\underline{C}^{P\cup Q}_{DE0}(X) = \underline{C}^P_{DE0}(X) \cup \varnothing = \underline{C}^P_{DE0}(X)$, no matter how many segmentations are performed, $\underline{C}_{DE0}(X)$ no longer change and $\underline{C}_{DE0}(X) = \underline{C}^P_{DE0}(X) = \underline{C}^{P\cup Q}_{DE0}(X)$.

In the set-valued decision system, $\underline{C}_{DE0}(X)$ increases with the increase of features. When enough features are added, $\underline{C}_{DE0}(X)$ remains stable and reaches the incremental limit of $\underline{C}_{DE0}(X)$. The main step is to continuously add features to the decision system, and use the incremental algorithm (Theorem 7) to quickly update $\underline{C}_{DE0}(X)$ until the limit of $\underline{C}_{DE0}(X)$ is reached (Theorem 8). When the limit of $\underline{C}_{DE0}(X)$ is reached, the added features and instances are the results of the collaborative reduction of features and instances. In addition, we introduce the Fisher score model, which greatly reduces the

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

blindness of adding features during incremental updates. The details are shown in Algorithm 1.

Algorithm 1: The collaborative reduction of features and instances based on the incremental updating of $\underline{C}_{DE0}(X)$ in semi-monolayer covering rough set (abbr. FSMCDE)

Input: $DS=(U,A\cup d,V,f)$ is a set-valued decision system, where $A=\{a_1,a_2,...,a_n\}$ and the division of decision d to U is $U/d=\{D_i|x\in U,|\vec{x}|_d=|\vec{y}|_d\}$.

Output: Feature subset FS after feature selection and instance set $LDE0$ after instance selection.

1) Calculate the Fisher score of each feature in $A=\{a_1,a_2,...,a_n\}$;

2) The features were sorted in descending order according to the Fisher score to obtain $A'=\{a'_1,a'_2,...,a'_n\}$;

3) $FS=\varnothing$, $LDE0=\varnothing$;

4) for $a'_q\in A'$ do

5) $FS=FS\cup\{a'_q\},DS'=(U,FS\cup d,V,f)$;

6) Calculate $Cell^Q,RC^Q$ *and* $RS(Cell^Q)$ according to Definition 3;

7) for $D_i\in U/d$ do

8) Calculate $\underline{C}^Q_{GC0}(D_i),\overline{C}^Q_{GC0}(D_i),\Delta LGC0^Q(D_i)$ *and* $\underline{C}^Q_{DE0}(D_i)$ according to Theorem 7;

9) if $\forall Di\in U/d$, satisfy the condition of $\underline{C}^Q_{GC0}(X)=\overline{C}^Q_{GC0}(X)$ according to Theorem 8;

10) $LDE0=\cup\underline{C}^Q_{DE0}(D_i)$

11) break;

12) Return: FS, $LDE0$

4. Experiment and analysis

In this section, we select ten datasets to test the effectiveness of the proposed FSMCDE algorithm, as shown in Table 1. All the experiments were performed on a computer with an Intel(R) Core (TM) i3-10100 CPU @ 3.60GHz and 32.0GB of RAM. The model in this paper was written by Scala 2.12 language and ran on IntelliJ IDEA. The data sets and source code in our paper have been uploaded to https://pan.baidu.com/s/1WNqopgeHfZyCxgFDJf3c1Q?pwd=o8um.

Table 1: The description of datasets

ID	Dataset	Abbreviation	Instances	Features	Classes
1	METABRIC	ME	2133	20000	6
2	Prostate	PR	102	10509	2
3	Breast2	BR	77	4869	2
4	dbworld_bodies	DB	64	4702	2
5	dbworld_bodies_stemmed	DBS	64	3721	2
6	Breast_Cancer 1	BC	168	2905	2
7	Musk1	MU	476	166	2
8	sonar	SO	208	60	2
9	Ionosphere	IO	351	34	2
10	wine	WI	178	13	3

In this paper, three sets of experiments are designed to verify the effectiveness of FSMCDE as follows. In the data tables, ODP represents no processing of the data set, the bolded and underlined data are the best-performing values, the symbol ' \ ' means that the corresponding result cannot be obtained.

4.1. Comparison with Feature Selection Algorithms

To verify the effectiveness of FSMCDE in feature selection, this section compares it with classical feature selection algorithms[14]. These algorithms are MI, Fisher Score, ReliefF, Chi-Square Score and F_score, where the number of features is set to 50. The feature selection results of the FSMCDE algorithm can be found in Table 6.

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

Table 2 and Table 3 show the classification results of feature selection algorithms on C4.5 and KNN, respectively. Combining Table 2 and Table 6, on the C4.5 classifier, FSMCDE can always achieve the highest accuracy (Acc) and F1 score with the fewest features. Combining Table 3 and Table 6, except for ME and PR, FSMCDE has the highest accuracy and F1 score on KNN. In summary, compared with classical feature selection algorithms, FSMCDE can effectively identify important features and achieve higher classification results with fewer features.

Table 2: Comparison of accuracy and F1 score of seven feature selection algorithms on C4.5 Classifier

Dataset	EI	ODP	ReliefF	CHI	F_score	MI	Fisher	FSMCDE
ME	Acc	70.48	61.744	61.818	71.269	69.77	70.534	**92.241**
	F1	70.52	61.715	61.815	71.277	69.818	70.504	**92.247**
PR	Acc	82.273	87.227	85.664	83.964	81.4	84.118	**89.8**
	F1	82.353	87.138	85.652	83.917	81.483	84.164	**89.811**
BR	Acc	57.5	68.054	67.643	71.5	71.375	69.839	**92.232**
	F1	56.85	67.549	66.943	71.074	71.444	69.759	**92.385**
DB	Acc	74.452	74.714	81.643	80.262	80.857	82.548	**86.333**
	F1	74.306	74.467	81.086	80.038	80.72	82.303	**86.265**
DBS	Acc	72.357	78.024	79.333	79.071	81	80.262	**85.233**
	F1	71.74	77.627	79.227	78.612	80.594	79.911	**85.267**
BC	Acc	75.408	70.783	72.044	77.195	70.724	76.68	**78.908**
	F1	75.143	70.928	71.729	77.112	70.671	76.528	**78.847**
MU	Acc	80.547	80.274	78.138	79.047	82.031	79.188	**95.23**
	F1	80.518	80.272	78.188	79.08	82.071	79.146	**95.209**
SO	Acc	72.048	77.781	76.017	75.705	74.912	75.479	**94.414**
	F1	72.012	77.881	75.964	75.606	74.833	75.396	**94.224**
IO	Acc	88.801	88.523	88.683	89.094	88.149	88.628	**96.51**
	F1	88.765	88.483	88.613	89.066	88.103	88.593	**96.504**
WI	Acc	93.778	93.105	93.108	92.915	93.578	93.598	**97.353**
	F1	93.765	93.069	93.05	92.932	93.523	93.575	**97.363**

Table 3: Comparison of accuracy and F1 score of seven feature selection algorithms on KNN Classifier

Dataset	EI	ODP	ReliefF	CHI	F_score	MI	Fisher	FSMCDE
ME	Acc	70.52	67.746	67.749	77.267	**78.603**	77.197	78.074
	F1	68.89	66.357	67.213	76.549	77.752	76.479	**77.961**
PR	Acc	79.545	82.291	88.818	**90.064**	88.745	**90.091**	89.9
	F1	79.722	82.178	88.812	**90.097**	88.839	**89.931**	90.018
BR	Acc	63.232	76.446	75.518	75.821	68.946	76.964	**89.964**
	F1	62.126	76.883	75.043	75.994	69.807	77.05	**89.747**
DB	Acc	55.429	88.714	87.31	87.214	88.595	86.357	**90.167**
	F1	41.623	88.548	87.333	87.202	88.439	85.744	**90.118**
DBS	Acc	56.19	92.071	90.524	90.119	91.214	90.238	**93.3**
	F1	43.598	92.147	90.677	90.295	91.376	90.38	**93.297**
BC	Acc	66.085	72.993	71.162	73.221	71.287	72.949	**81.667**
	F1	53.158	67.961	66.864	69.777	64.836	69.405	**79.552**
MU	Acc	84.474	82.44	70.913	78.907	84.542	79.039	**92.285**
	F1	84.538	82.401	70.849	78.827	84.498	78.97	**92.304**
SO	Acc	70.348	69.945	70.957	71.324	71.879	70.402	**91.052**
	F1	70.13	69.546	70.855	71.304	71.777	70.375	**88.554**
IO	Acc	82.585	82.596	82.525	82.485	82.108	82.511	**87.461**
	F1	81.141	81.141	81.026	80.928	80.605	81.013	**87.56**
WI	Acc	68.31	68.552	68.382	68.16	68.493	68.454	**75.33**
	F1	67.763	68.163	68.308	67.807	67.964	67.942	**70.372**

4.2. Comparison with Instance Selection Algorithms

To further verify the effectiveness of FSMCDE in the instance selection, we compare it with the fuzzy rough prototype selection (FRPS)[7]. It is determined to use lower (FRPSI) and upper (FRPSII) approximation membership as a quality measure to select instances.

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

In Table 4 and Table 5, FSMCDE can always achieve the best classification results on C4.5 and KNN classifiers compared with FRPSI and FRPSII. In terms of the number of instances, it is clear that FSMCDE is more suitable for processing high-dimensional data sets. FSMCDE can effectively obtain reasonable instance selection results, while FRPSI and FRPSII will get more extreme instance selection results, such as filtering too many instances, or not filtering any instances. Therefore, FSMCDE can perform effective instance selection to obtain high-quality instances and improve classification results.

Table 4: The number of instances selected by the three instance selection algorithms and the accuracy and F1 score on the C4.5 classifier

Dataset	FRPSI			FRPSII			FSMCDE		
	Instances	Acc	F1	Instances	Acc	F1	Instances	Acc	F1
ME	\	\	\	\	\	\	2025	**92.24**	**92.25**
PR	96	81.02	80.88	102	82.96	82.9	100	**89.8**	**89.81**
BR	16	67	67.67	45	64.25	64.54	74	**92.23**	**92.39**
DB	16	49.5	49.33	64	70.91	70.9	60	**86.33**	**86.27**
DBS	19	52	52	64	75.14	75.07	55	**85.23**	**85.27**
BC	45	65	64.12	149	67.13	67.23	157	**78.91**	**78.85**
MU	309	84.34	84.36	476	81.64	81.64	403	**95.23**	**95.21**
SO	200	73.25	73.02	208	74.21	74.15	206	**94.41**	**94.22**
IO	174	92.07	92.05	351	89	88.98	347	**96.51**	**96.5**
WI	143	92.88	92.82	178	93.75	93.71	174	**97.35**	**97.36**

Table 5: The number of instances selected by the three instance selection algorithms and the accuracy and F1 score on the KNN classifier

Dataset	FRPSI			FRPSII			FSMCDE		
	Instances	Acc	F1	Instances	Acc	F1	Instances	Acc	F1
ME	\	\	\	\	\	\	2025	**78.07**	**77.96**
PR	96	82.12	82.12	102	78.8	78.98	100	**89.9**	**90.02**
BR	16	69.5	66	45	73.3	67.22	74	**89.96**	**89.75**
DB	16	50	49	64	55.48	41.51	60	**90.17**	**90.12**
DBS	19	72	66.33	64	56.55	44.51	55	**93.3**	**93.3**
BC	45	62.3	49.97	149	69.87	58.1	157	**81.67**	**79.55**
MU	309	88.57	88.67	476	84.37	84.46	403	**92.29**	**92.3**
SO	200	70.8	70.51	208	70.8	70.51	206	**91.05**	**88.55**
IO	174	69.68	65.4	351	82.1	80.48	347	**87.46**	**87.56**
WI	143	71.31	71.47	178	69.3	68.86	174	**75.33**	**70.37**

4.3. Comparison with Feature Selection + Instance Selection Algorithms

Table 6: The number of instances, the number of features and the accuracy, F1 score on the C4.5 classifier of the three features + instance selection algorithms

Dataset	FRPSII-FISC			FISC-FRPSII			FSMCDE		
	(Ins, Fea)	Acc	F1	(Ins, Fea)	Acc	F1	(Ins, Fea)	Acc	F1
ME	\	\	\	(1253,50)	79.63	79.71	(2025,42)	**92.24**	**92.25**
PR	(102, 50)	83	83.11	(101, 50)	84.47	84.44	(100,10)	**89.8**	**89.81**
BR	(45, 50)	79.55	79.37	(35, 50)	81.17	79.96	(74,6)	**92.23**	**92.39**
DB	(64, 50)	80.67	80.62	(61, 50)	87.07	87.16	(60,24)	**86.33**	**86.27**
DBS	(64, 50)	80.57	80.24	(58, 50)	89.27	89.44	(55,21)	**85.23**	**85.27**
BC	(149, 50)	73.61	73.38	(146, 50)	76.32	76.48	(157,28)	**78.91**	**78.85**
MU	(476, 50)	78.85	78.84	(458, 50)	78.52	78.54	(403,40)	**95.23**	**95.21**
SO	(208, 50)	74.61	74.54	(208, 50)	75.1	75	(206,13)	**94.41**	**94.22**
IO	(351, 34)	88.92	88.91	(351, 34)	88.44	88.4	(347,24)	**96.51**	**96.5**
WI	(178, 13)	92.93	92.89	(178, 13)	93.11	93.1	(174,9)	**97.35**	**97.36**

Finally, this paper compares FSMCDE with feature + instance algorithm to show the superiority of FSMCDE in feature and instance collaborative reduction. We use the Fisher score and FRPS II with better classification results to combine, and the number of features is set to 50. Table 6 shows the number

ISSN 2616-5775 Vol. 7, Issue 3: 13-20, DOI: 10.25236/AJCIS.2024.070302

of features and instances selected by the algorithms. The format is (number of instances, number of features), and we abbreviate it as (Ins, Fea). Fisher score is abbreviated as FISC.

According to Table 6, FSMCDE has the highest classification results on all datasets. The accuracy and F1 score of FSMCDE were 7.2 % and 7.3 % higher than the second place, respectively. FSMCDE can always show stability on high-dimensional data sets, and the two comparison algorithms will frequently fail to filter out any instance or filter out too many instances. Therefore, FSMCDE can better identify high-resolution features and noise instances, efficiently perform collaborative reduction, improve data quality, and effectively improve classification performance.

5. Conclusions

Based on the incremental update theory of the DE0 lower approximation set in semi-monolayer covering rough set, this paper proposes the feature and instance collaborative reduction algorithm (FSMCDE). Firstly, the incremental update theory and incremental limit of DE0 lower approximation set are designed. Secondly, features are continuously added to the set-valued system, and the incremental algorithm is used to quickly update the lower approximation set until it reaches the limit, thereby completing the collaborative reduction of features and instances. Finally, experiments show that FSMCDE can efficiently perform collaborative reduction and improve classification performance, which is basically better than all comparison algorithms. In addition, this paper does not consider the imbalance problem of the data set, and the method of dealing with the class imbalance problem will be fruitful. We will address these issues as part of our future work.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (No. 61972134).

References

[1] Idri A, Benhar H, Fernández-Alemán J L, et al. A systematic map of medical data preprocessing in knowledge discovery[J]. Computer Methods and Programs in Biomedicine. 2018, 162: 69-85.
[2] Buza K, Nanopoulos A, Schmidt-Thieme L. INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification[C]. In Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part II 15, pp. 149-160. Springer Berlin Heidelberg, 2011.
[3] Cai J, Luo J, Wang S, et al. Feature selection in machine learning: A new perspective[J]. Neurocomputing. 2018, 300: 70-79.
[4] Pawlak Z, Skowron A. Rudiments of rough sets[J]. Information Sciences. 2007, 177(1): 3-27.
[5] Sun L, Zhang X, Qian Y, et al. Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification[J]. Information Sciences. 2019, 502: 18-41.
[6] R. J, M. A, N. M P. Effective instance selection using the fuzzy-rough lower approximation[C]. 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2019.
[7] Verbiest N, Cornelis C, Herrera F. FRPS: A Fuzzy Rough Prototype Selection method[J]. Pattern Recognition. 2013, 46(10): 2770-2782.
[8] Anaraki J, Samet S, Lee J, et al. SUFFUSE: Simultaneous Fuzzy-Rough Feature- Sample Selection[J]. Journal of Advances in Information Technology. 2015, 6: 103-110.
[9] Derrac J, Cornelis C, García S, et al. Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection[J]. Information Sciences. 2012, 186(1): 73-92.
[10] Shu W, Shen H. Updating attribute reduction in incomplete decision systems with the variation of attribute set[J]. International Journal of Approximate Reasoning. 2014, 55(3): 867-884.
[11] Wu Z, Chen N, Gao Y. Semi-monolayer cover rough set: Concept, property and granular algorithm [J]. Information Sciences. 2018, 456: 97-112.
[12] Wu Z, Wang H, Chen N, et al. Semi-monolayer covering rough set on set-valued information systems and its efficient computation[J]. International Journal of Approximate Reasoning. 2021, 130: 83-106.
[13] Guan Y, Wang H. Set-valued information systems[J]. Information Sciences. 2006, 176(17): 2507-2525.
[14] Li J, Cheng K, Wang S, et al. Feature Selection: A Data Perspective[J]. ACM Comput. Surv. 2017, 50(6): 94.