Online multi-label streaming feature selection by label causal relationships and multigranulation fuzzy implication information

1. Preliminaries

1.1. Multigranulation fuzzy rough sets

$FIS = \langle U, AS \rangle$ is a fuzzy information system where $U = \{u_1, u_2, \ldots, u_n\}$ is a universe, and $AS = \{f_1, f_2, \ldots, f_m\}$ is a non-empty finite set of attributes.

Definition 1 ([1]). For $\forall X \subseteq U$, $P_1, P_2, \ldots, P_k \subseteq AS$, $R_{P_i}(1 \leq i \leq k)$ is a family of fuzzy relations, the optimistic multigranulation lower and upper approximations are denoted by $\sum_{i=1}^{k} \underline{R_{P_i}^{O}}$ and $\sum_{i=1}^{k} \overline{R_{P_i}^{O}}$, respectively,

$$\sum_{i=1}^{k} \underline{R_{P_i}^{O}} X(x) = \vee_{i=1}^{k} \wedge_{y \in U} ((1 - R_{P_i}(x,y)) \vee X(y)), \tag{1}$$

$$\sum_{i=1}^{k} \overline{R_{P_i}^{O}} X(x) = \wedge_{i=1}^{k} \vee_{y \in U} (R_{P_i}(x,y) \wedge X(y)). \tag{2}$$

Definition 2 ([1]). For $\forall X \subseteq U$, $P_1, P_2, \ldots, P_k \subseteq AS$, $R_{P_i}(1 \leq i \leq k)$ is a family of fuzzy relations, the pessimistic multigranulation lower and upper approximations are denoted by $\sum_{i=1}^{k} \underline{R_{P_i}^{P}}$ and $\sum_{i=1}^{k} \overline{R_{P_i}^{P}}$, respectively,

$$\sum_{i=1}^{k} \underline{R_{P_i}^{P}} X(x) = \wedge_{i=1}^{k} \wedge_{y \in U} ((1 - R_{P_i}(x,y)) \vee X(y)), \tag{3}$$

$$\sum_{i=1}^{k} \overline{R_{P_i}^{P}} X(x) = \vee_{i=1}^{k} \vee_{y \in U} (R_{P_i}(x,y) \wedge X(y)). \tag{4}$$

1.2. Bayesian networks

The concept of Markov Blanket (MB) originates from causal Bayesian networks, where a Bayesian network that satisfies faithfulness, has a unique set of MB for each variable [2], and this Bayesian network can be used to describe the causal structure in a collection of variables.

Definition 3 (Bayesian Network [2]). A Bayesian network BN is defined as $\langle \mathcal{V}, G, P \rangle$, where $\mathcal{V}$ represents the set of variables, G is a directed acyclic graph (DAG), and P represents the joint probability distribution over the variables $\mathcal{V}$ in G, and the BN satisfies the Markov condition: for any variable X, given its parent variables, every non-descendant variable in the network is independent of X.

Bayesian networks are models used to describe causal relationships among a collection of variables. For a pair of parent-child variables in DAG, the child is directly influenced by the parent and the parent is a direct cause of the child. A fundamental assumption in Bayesian networks is faithfulness, which can be described as follows:

Definition 4 (Faithfulness Assumption [2]). Given a Bayesian network $BN = \langle \mathcal{V}, G, P \rangle$, G is faithful to P if and only if all conditional independence relations in P are determined by G and the Markov condition. Similarly, P is faithful to G if and only if a subgraph of G exists that is faithful to P.

When the faithfulness assumption holds, the Markov Blanket provides comprehensive information about the class labels [3]. Therefore, the relationship between labels can be found in the MB of the faithful *BN*. The concept of MB is based on the faithful *BN*, which has the following definition:

Definition 5 (Markov Blanket (MB) [2]). In a *BN* that meets faithfulness, the Markov blanket of a variable includes the parents, children, and spouses variables of this variable.

2. Multigranulation fuzzy implication information

Definition 6. *The optimistic and pessimistic multigranulation fuzzy implication information is defined as:*

$$\widetilde{MFI}^{O}(\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|\cup_{s=1}^{k}[x_i]_{P_s}|}{n}\right), \tag{5}$$

$$\widetilde{MFI}^{P}(\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|\cap_{s=1}^{k}[x_i]_{P_s}|}{n}\right), \tag{6}$$

where $[x_i]_{P_s}$ denotes the fuzzy similarity class of x_i induced by R_{P_s}, and $|\cdot|$ represents the cardinality of the fuzzy similarity class.

Remark 1. When $P_1 = P_2 = \cdots = P_k$, Eq.(5) and Eq.(6) reduces to

$$\widetilde{MFI}^{O}(\sum_{s=1}^{k} P_s) = \widetilde{MFI}^{P}(\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|[x_i]_{P_s}|}{n}\right). \tag{7}$$

This also means that multigranulation fuzzy implication information reduces to fuzzy implication granularity information [4].

Proposition 1. Suppose $P_1, P_2, \ldots, P_k \subseteq C$. The following properties are satisfied.

1. $\widetilde{MFI}^{O}(\sum_{s=1}^{k} P_s) \leq \widetilde{MFI}^{P}(\sum_{s=1}^{k} P_s)$;

2. When $B \subseteq E \subseteq C$, we can get $\widetilde{MFI}^{O}(\sum_{P_s \in B} P_s) \geq \widetilde{MFI}^{O}(\sum_{P_s \in E} P_s)$ and $\widetilde{MFI}^{P}(\sum_{P_s \in B} P_s) \leq \widetilde{MFI}^{P}(\sum_{P_s \in E} P_s)$.

Proof.

1. Since $|\cap_{s=1}^{k}[x_i]_{P_s}| \leq |\cup_{s=1}^{k}[x_i]_{P_s}|$. Therefore, $\widetilde{MFI}^{O}(\sum_{s=1}^{k} P_s) \leq \widetilde{MFI}^{P}(\sum_{s=1}^{k} P_s)$;

2. If $B \subseteq E$, then $|\cup_{P_s \in B}[x_i]_{P_s}| \leq |\cup_{P_s \in E}[x_i]_{P_s}|$ and $|\cap_{P_s \in B}[x_i]_{P_s}| \geq |\cap_{P_s \in E}[x_i]_{P_s}|$. Hence, $\widetilde{MFI}^{O}(\sum_{P_s \in B} P_s) \geq \widetilde{MFI}^{O}(\sum_{P_s \in E} P_s)$ and $\widetilde{MFI}^{P}(\sum_{P_s \in B} P_s) \leq \widetilde{MFI}^{P}(\sum_{P_s \in E} P_s)$.

The second property indicates that the multigranulation fuzzy implication information varies monotonically with the size of the covering family. It is expected that the addition of new coverings will enhance the system's ability to distinguish between elements.

Definition 7. *Given a label d, the optimistic and pessimistic multigranulation fuzzy conditional implication information of d relative to $P_1, P_2, \ldots, P_k$ is defined as follows:*

$$\widetilde{MFI}^{O}(d|\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|[x_i]_d \cup [x_i]_{P'}^{O}|}{n} - \frac{|[x_i]_d|}{n}\right), \tag{8}$$

$$\widetilde{MFI^P}(d|\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|[x_i]_d \cup [x_i]_{P'}^{P}|}{n} - \frac{|[x_i]_d|}{n}\right), \quad (9)$$

where $[x_i]_{P'}^{O} = \cup_{s=1}^{k}[x_i]_{P_s}$, $[x_i]_{P'}^{P} = \cap_{s=1}^{k}[x_i]_{P_s}$.

Proposition 2. Suppose $P_1, P_2, \ldots, P_k \subseteq C$, if $\cup_{s=1}^{k}[x_i]_{P_s} \subseteq [x_i]_d$, $\cap_{s=1}^{k}[x_i]_{P_s} \subseteq [x_i]_d$, then $\widetilde{MFI^O}(d|\sum_{s=1}^{k} P_s) = 0$ and $\widetilde{MFI^P}(d|\sum_{s=1}^{k} P_s) = 0$.

Proof. Since $\cup_{s=1}^{k}[x_i]_{P_s} \subseteq [x_i]_d$, $\cap_{s=1}^{k}[x_i]_{P_s} \subseteq [x_i]_d$, so $|[x_i]_d \cup (\cup_{s=1}^{k}[x_i]_{P_s})| = |[x_i]_d|$, $|[x_i]_d \cup (\cap_{s=1}^{k}[x_i]_{P_s})| = |[x_i]_d|$. Therefore $\widetilde{MFI^O}(d|\sum_{s=1}^{k} P_s) = 0$ and $\widetilde{MFI^P}(d|\sum_{s=1}^{k} P_s) = 0$.

Proposition 3. Suppose $P_1 \cup P_2 \cdots \cup P_k = B$, $\tilde{P}_1 \cup \tilde{P}_2 \cdots \cup \tilde{P}_g = E$, if $B \subseteq E \subseteq C$, then $\widetilde{MFI^P}(d|\sum_{s=1}^{g} \tilde{P}_s) \leq \widetilde{MFI^P}(d|\sum_{s=1}^{k} P_s)$.

Proof. Since $P_1 \cup P_2 \cdots \cup P_k = B$, $\tilde{P}_1 \cup \tilde{P}_2 \cdots \cup \tilde{P}_g = E$, $B \subseteq E \subseteq C$, so $\cap_{s=1}^{g}[x_i]_{\tilde{P}_s} \subseteq \cap_{s=1}^{k}[x_i]_{P_s}$, then $|[x_i]_d \cup (\cap_{s=1}^{g}[x_i]_{\tilde{P}_s})| \leq |[x_i]_d \cup (\cap_{s=1}^{k}[x_i]_{P_s})|$. Therefore $\widetilde{MFI^P}(d|\sum_{s=1}^{g} \tilde{P}_s) \leq \widetilde{MFI^P}(d|\sum_{s=1}^{k} P_s)$.

Definition 8. *The optimistic and pessimistic multigranulation fuzzy mutual implication information of* $P_1, P_2, \ldots, P_k$ *and* d *is defined as follows:*

$$\widetilde{MFMI^O}(\sum_{s=1}^{k} P_s; d) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|[x_i]_{P'}^{O} \cup [x_i]_d|}{n}\right), \quad (10)$$

$$\widetilde{MFMI^P}(\sum_{s=1}^{k} P_s; d) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|[x_i]_{P'}^{P} \cup [x_i]_d|}{n}\right). \quad (11)$$

Proposition 4. Suppose $P_1, P_2, \ldots, P_k \subseteq C$, then $0 \leq \widetilde{MFMI^O}(\sum_{s=1}^{k} P_s; d) < 1$ and $0 \leq \widetilde{MFMI^P}(\sum_{s=1}^{k} P_s; d) < 1$.

Proof. Since $0 < |(\cup_{s=1}^{k}[x_i]_{P_s}) \cup [x_i]_d| \leq |U|$, $0 < |(\cap_{s=1}^{k}[x_i]_{P_s}) \cup [x_i]_d| \leq |U|$, and $|U| = n$, then $0 < \frac{|(\cup_{s=1}^{k}[x_i]_{P_s}) \cup [x_i]_d|}{n} \leq 1$ and $0 < \frac{|(\cap_{s=1}^{k}[x_i]_{P_s}) \cup [x_i]_d|}{n} \leq 1$. Therefore $0 \leq \widetilde{MFMI^O}(\sum_{s=1}^{k} P_s; d) < 1$ and $0 \leq \widetilde{MFMI^P}(\sum_{s=1}^{k} P_s; d) < 1$.

Definition 9. *Given* $f \in C$, *and the optimistic and pessimistic multigranulation fuzzy conditional mutual implication information of* d *and* f *relative to* $P_1, P_2, \ldots, P_k$ *is defined as follows:*

$$\widetilde{MFMI^O}(d; f|\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|[x_i]_f \cup [x_i]_{P'}^{O}|}{n} - \frac{|[x_i]_f \cup [x_i]_{P' \cup d}^{O}|}{n}\right), \quad (12)$$

$$\widetilde{MFMI^P}(d; f|\sum_{s=1}^{k} P_s) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|[x_i]_f \cup [x_i]_{P'}^{P}|}{n} - \frac{|[x_i]_f \cup [x_i]_{P' \cup d}^{P}|}{n}\right), \quad (13)$$

where $[x_i]_{P' \cup d}^{O} = [x_i]_{P'}^{O} \cap [x_i]_d$, $[x_i]_{P' \cup d}^{P} = [x_i]_{P'}^{P} \cap [x_i]_d$.

Proposition 5. Suppose $P_1, P_2, \ldots, P_k \subseteq C$, then $0 \leq \widetilde{MFMI^O}(d; f|\sum_{s=1}^{k} P_s) < 1$ and $0 \leq \widetilde{MFMI^P}(d; f|\sum_{s=1}^{k} P_s) < 1$.

Proof. Since $((\cup_{s=1}^{k}[x_i]_{P_s}) \cap [x_i]_d) \subseteq (\cup_{s=1}^{k}[x_i]_{P_s})$ and $((\cap_{s=1}^{k}[x_i]_{P_s}) \cap [x_i]_d) \subseteq (\cap_{s=1}^{k}[x_i]_{P_s})$, then $|[x_i]_f \cup (\cup_{s=1}^{k}[x_i]_{P_s})| \geq |[x_i]_f \cup ((\cup_{s=1}^{k}[x_i]_{P_s}) \cap [x_i]_d)|$ and $|[x_i]_f \cup (\cap_{s=1}^{k}[x_i]_{P_s})| \geq |[x_i]_f \cup ((\cap_{s=1}^{k}[x_i]_{P_s}) \cap [x_i]_d)|$. Therefore $0 \leq \widetilde{MFMI}^{O}(d; f|\sum_{s=1}^{k} P_s)$ and $0 \leq \widetilde{MFMI}^{P}(d; f|\sum_{s=1}^{k} P_s)$. Equality holds when $(\cup_{s=1}^{k}[x_i]_{P_s}) \subseteq [x_i]_d$ and when $(\cap_{s=1}^{k}[x_i]_{P_s}) \subseteq [x_i]_d$, respectively. $\widetilde{MFMI}^{O}(d; f|\sum_{s=1}^{k} P_s) < 1$ and $\widetilde{MFMI}^{P}(d; f|\sum_{s=1}^{k} P_s) < 1$ can be easily proved. Therefore $0 \leq \widetilde{MFMI}^{O}(d; f|\sum_{s=1}^{k} P_s) < 1$ and $0 \leq \widetilde{MFMI}^{P}(d; f|\sum_{s=1}^{k} P_s) < 1$.

3. Experimental

3.1. Acquisition of fuzzy similarity relation

A multi-label dataset can be viewed as a multi-label decision table (U, C, L), $\forall u_i, u_j \in U$, $E \in C$, the similarity relation between u_i and u_j on E is defined as:

$$R_E(u_i, u_j) = \exp\left(-\frac{\|u_i - u_j\|^2}{2 \cdot \sigma^2}\right), \tag{14}$$

where σ represents the standard deviation and $\|\bullet\|$ denotes Euclidean distance.

Table 1: Description of datasets.

No.	Dataset	Samples	Train	Test	Features	Labels	Dom	Card	Dens
1	Birds	645	322	323	258	19	Audio	1.014	0.053
2	Business	5000	2000	3000	438	30	Text	1.588	0.053
3	Computers	5000	2000	3000	681	33	Text	1.508	0.046
4	Genbase	662	463	199	1186	27	Biology	1.252	0.046
5	Medical	978	333	645	1449	45	Text	1.245	0.028
6	Society	5000	2000	3000	636	27	Text	1.692	0.063
7	Yeast	2417	1500	917	103	14	Biology	4.237	0.303
8	Cal500	502	350	152	68	174	Music	26.044	0.15
9	Social	5000	2000	3000	1047	40	Text	1.283	0.033
10	Scene	2407	1211	1196	294	6	Image	1.074	0.179

3.2. Multi-label benchmark datasets and experimental settings

To evaluate the performance of the proposed algorithm, we conduct experiments on 10 benchmark datasets drawn from diverse application domains. Specifically, datasets such as Business, Computers, Medical, Society, and Social come from the text domain and are used for webpage classification, while Birds, Genbase, Yeast, and Scene are from the audio, biology, and image domains, respectively. The streaming feature scenario is simulated using the datasets listed in Table 1. All experiments are carried out on a PC running Windows, equipped with a 3.7 GHz CPU and 16 GB of RAM. The algorithm is implemented in MATLAB.

3.3. Experimental results

We use four commonly used evaluation metrics for multi-label learning [5] to assess the performance of multi-label algorithms. Larger values of AP represent better classification performance and smaller values of OE, RL, and CV represent better classification performance. Lastly, the ML-KNN (K = 10) classifier [6] is used to evaluate the performance of all multi-label feature selection algorithms. The detailed experimental results are presented in Tables 2-5.

3.4. Parameter analysis

To verify the optimality and stability of parameter δ, we conducted a parameter sensitivity analysis, as shown in Figure 1. We varied δ within the range [0.01,0.1] with a step size of 0.01 and recorded the performance. As shown in the figure, our chosen thresholds (δ=0.03 for text data, i.e., Business and Society; δ=0.05, i.e., Scene, Cal500 and Birds) fall within the performance plateau, and are optimal or near-optimal across all datasets. This comprehensive sensitivity analysis validates the robustness and rationality of our parameter settings.

Table 2: Comparative evaluation among algorithms in terms of $AP(\uparrow)$.

Dataset	OMSFS_{LC}	OMGFS	OM-LEFSD	NRFSFN	SRFS	SRLG-LMA	OM-LCMFI
Birds	0.6985	0.6985	0.6875	0.6903	0.6445	0.6846	**0.7020**
Business	0.8756	0.8759	0.8690	0.8756	0.8708	**0.8768**	0.8765
Computers	**0.6423**	0.6246	0.6328	0.6342	0.6231	0.6387	0.6414
Genbase	0.9794	0.7114	0.9889	0.6840	0.9898	0.9899	**0.9914**
Medical	0.5479	0.6667	0.7434	0.4180	0.6852	0.7384	**0.7900**
Society	**0.5981**	0.5945	0.5877	0.5906	0.5619	0.5862	0.5972
Yeast	0.7510	0.7536	0.7521	0.7514	0.7344	0.7527	**0.7571**
Cal500	0.4927	0.4961	0.4910	0.4932	0.4966	0.4923	**0.4975**
Social	0.7064	0.7133	0.7008	0.7133	0.6978	0.7122	**0.7181**
Scene	0.8396	0.8456	0.8385	0.8261	0.7410	0.8076	**0.8461**
Average	0.7132	0.6980	0.7292	0.6677	0.7045	0.7279	**0.7417**

Table 3: Comparative evaluation among algorithms in terms of $OE(\downarrow)$.

Dataset	OMSFS_{LC}	OMGFS	OM-LEFSD	NRFSFN	SRFS	SRLG-LMA	OM-LCMFI
Birds	0.3653	0.3715	0.4025	0.3994	0.4613	0.3932	**0.3591**
Business	0.1240	0.1233	0.1333	0.1253	0.1303	**0.1220**	0.1227
Computers	0.4300	0.4560	0.4377	0.4380	0.4543	0.4393	**0.4287**
Genbase	0.0201	0.4121	0.0101	0.4523	**0.0050**	**0.0050**	**0.0050**
Medical	0.5829	0.4558	0.3178	0.7147	0.3891	0.3426	**0.2589**
Society	**0.4427**	0.4503	0.4620	0.4533	0.4930	0.4587	0.4453
Yeast	0.2465	0.2399	0.2465	0.2345	0.2508	0.2388	**0.2334**
Cal500	0.1053	0.1053	**0.0987**	0.1053	**0.0987**	0.1053	**0.0987**
Social	0.3863	0.3727	0.3977	0.3703	0.4007	0.3747	**0.3673**
Scene	0.2676	**0.2542**	0.2634	0.2818	0.4030	0.3211	**0.2542**
Average	0.2971	0.3241	0.2770	0.3575	0.3086	0.2801	**0.2573**

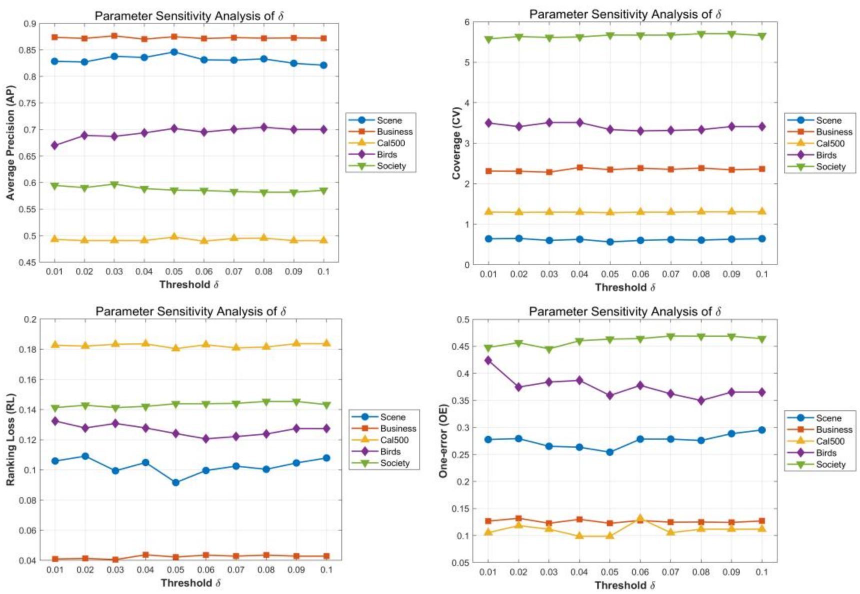

Figure 1: Parameter analysis

Table 4: Comparative evaluation among algorithms in terms of $CV(\downarrow)$.

Dataset	OMSFS_{LC}	OMGFS	OM-LEFSD	NRFSFN	SRFS	SRLG-LMA	OM-LCMFI
Birds	3.5015	3.4025	**3.2663**	3.3003	3.8607	3.3498	3.3375
Business	2.3413	2.2963	2.3317	2.2997	2.3903	**2.2840**	2.2853
Computers	**4.3413**	4.4320	4.5337	4.4973	4.4923	4.3763	4.3527
Genbase	0.6482	1.7337	0.5829	2.2261	0.5879	0.5779	**0.5779**
Medical	5.0961	4.0481	3.6295	6.7814	4.2202	3.4155	**3.1426**
Society	5.6483	5.6580	5.6410	5.6550	5.9140	5.7167	**5.6120**
Yeast	6.4733	6.4569	6.4406	6.5573	6.6619	6.4995	**6.3697**
Cal500	128.6118	128.6842	129.1250	128.4211	129.2500	130.6053	**128.3553**
Social	3.4667	**3.3937**	3.5160	3.4740	3.5160	3.4390	3.4000
Scene	0.5669	0.5686	0.6020	0.6388	1.0159	0.6831	**0.5619**
Average	16.0696	16.0674	15.9669	16.3851	16.1909	16.0947	**15.7995**

Table 5: Comparative evaluation among algorithms in terms of $RL(\downarrow)$.

Dataset	OMSFS_{LC}	OMGFS	OM-LEFSD	NRFSFN	SRFS	SRLG-LMA	OM-LCMFI
Birds	0.1289	0.1286	0.1247	**0.1235**	0.1470	0.1267	0.1241
Business	0.0416	0.0411	0.0428	0.0411	0.0431	0.0407	**0.0405**
Computers	0.0906	0.0922	0.0944	0.0936	0.0938	0.0907	0.0901
Genbase	0.0099	0.0513	0.0066	0.0667	0.0081	0.0081	**0.0064**
Medical	0.0941	0.0711	0.0629	0.1318	0.0746	0.0575	**0.0516**
Society	0.1429	0.1426	0.1432	0.1429	0.1517	0.1451	**0.1413**
Yeast	0.1771	0.1743	0.1767	0.1812	0.1898	0.1773	**0.1724**
Cal500	0.1825	0.1809	0.1836	0.1808	0.1810	0.1827	**0.1804**
Social	0.0659	**0.0634**	0.0667	0.0650	0.0667	0.0647	0.0637
Scene	0.0930	0.0926	0.0991	0.1074	0.1823	0.1156	**0.0916**
Average	0.1027	0.1038	0.1001	0.1134	0.1138	0.1009	**0.0962**

References

[1] B. Sun, W. Ma, Y. Qian, Multigranulation fuzzy rough set over two universes and its application to decision making, Knowledge-Based Systems 123 (2017) 61–74.

[2] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, 2014.

[3] T. Gao, Q. Ji, Local causal discovery of direct causes and effects, Advances in Neural Information Processing Systems 28 (2015) 2512–2520.

[4] J. Dai, Z. Zhu, X. Zou, Fuzzy rough attribute reduction based on fuzzy implication granularity information, IEEE Transactions on Fuzzy Systems 32 (6) (2024) 3741–3752.

[5] W. Qian, J. Huang, F. Xu, W. Shu, W. Ding, A survey on multi-label feature selection from perspectives of label fusion, Information Fusion 100 (2023) 101948.

[6] M.-L. Zhang, Z.-H. Zhou, Ml-knn: A lazy learning approach to multi-label learning, Pattern Recognition 40 (7) (2007) 2038–2048.