Journal of Optimization of Soft Computing (JOSC)
Vol. 3, Issue 1, pp: (22-31), Summer-2025
Journal homepage: https://sanad.iau.ir/journal/josc

Paper Type (Research paper)

An Efficient Approach for Multi-Label Streaming Feature Selection

Azar Rafie[1], Parham Moradi[2*]

1. Department of Computer Engineering, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran
2. School of engineering, RMIT University Melbourne, Australia

Article Info

Article History:
Received: 2025/07/25
Revised: 2025/08/30
Accepted: 2025/09/14

DOI: josc.2025.202507241212912

Keywords:
Streaming multi-label data, feature selection, mutual information, redundancy, Relevancy.

**Corresponding Author's Email: p.moradi@uok.ac.ir*

Abstract

With the rapid growth of multi-label streaming data, efficient feature selection becomes a critical challenge. Traditional methods often struggle to handle the dynamic nature of continuously arriving data. This paper introduces OSM-MI, a novel online feature selection method designed for multi-label streaming datasets. OSM-MI uses mutual information to dynamically select features, minimizing redundancy and maximizing relevance. The method is compared with existing algorithms, including OM-NRS, OMGFS, and MUCO, across several datasets such as Yeast, Medical, Scene, Enron, and others. Experimental results show that OSM-MI outperforms the other methods in terms of accuracy, precision, and efficiency, while also maintaining lower execution times. Statistical significance is confirmed through the Wilcoxon test, demonstrating OSM-MI's robustness for real-time multi-label classification. This work provides an efficient, scalable solution for feature selection in streaming environments.

1. Introduction

With the rapid growth of online data such as images, videos, user comments, and tweets, there is a critical need for scalable classification systems to manage and search this content. Data mining and machine learning algorithms lose their effectiveness when dealing with large-scale data, and feature selection can address this issue. This process enhances algorithm performance by reducing data dimensions and selecting relevant features [1, 2]. Feature selection also helps reduce memory requirements, modeling time, and improves the performance of predictive algorithms [3, 4]. The goal of feature selection is to choose a subset of features relevant to class labels in order to build an efficient predictive model. Feature selection leads to reduced memory requirements for storage, decreased modeling and training time in machine learning algorithms, improved performance of predictive algorithms, better data understanding, among other benefits [5].

Traditional feature selection algorithms assume that all features are available before the feature selection process begins. However, in real-world scenarios, features are gradually and dynamically added to the data. For instance, in image analysis and satellite data, features are continuously added to the training data [۷ ,۶]. Therefore, online feature selection becomes essential [۹ ,۸]. Online feature selection algorithms are divided into two categories: the first adds features incrementally, while the second increases samples online.

Additionally, features can be produced in groups, which require specific algorithms [١٠ ,١١, 12] .
In multi-label data, online feature selection must be able to identify features relevant to all labels. Various methods for feature selection in multi-label data have been proposed, including approaches based on mutual information and redundancy analysis [١٣, ١٤] . These methods help select effective features and reduce redundancy. The main focus in these methods is to select features that distinguish objects from their surrounding environment. Since the background and foreground are constantly changing, the use of an online and adaptive algorithm for object identification is very effective. Additionally, in content-based image retrieval [١٥], the online learning process must address a core issue, which is identifying features that better represent the current query concept. To solve this issue, this paper proposes a method for feature selection in multi-label training data with feature streaming.
In this study, the novelty lies in designing an online feature selection framework that simultaneously addresses the challenges of streaming features and multi-label data, which existing methods often treat separately. Unlike conventional approaches that either focus only on incremental features or only on label correlations, our method integrates both aspects to capture more representative and less redundant features. The main contributions of this work are threefold: (1) we introduce a dynamic mechanism for selecting features in real-time under streaming conditions, (2) we incorporate multi-label dependency modeling to enhance relevance across multiple classes, and (3) we demonstrate through experimental validation that our approach achieves superior performance compared to state-of-the-art methods in terms of accuracy, scalability, and adaptability. These contributions highlight the significance of the proposed method and establish its practical relevance for large-scale, real-world applications.

2. Related Work

Based on the premise that features or training samples are gradually added to the dataset over time, there are different online feature selection algorithms. In datasets where features are gradually added over time, feature stream-based selection algorithms are used. When samples are added over time, sample stream-based feature selection algorithms are applied. When both features and samples are gradually added to the dataset, these algorithms are referred to as feature and sample stream-based feature selection algorithms [16].
In individual online feature selection methods, it is assumed that features are added to the dataset one by one. Perkins and Taylor [١٠] introduced a graph-based method for online feature selection, which relies on error gradient reduction. Zhu and colleagues [١١] proposed two regression-based algorithms, Information-Investing and Alpha-Investing, for online feature selection. Wu and colleagues [١٧] introduced the OSFS and Fast-OSFS algorithms. Yu and colleagues [٨] proposed the SAOLA feature selection algorithm. These algorithms serve as the foundation for various online feature selection methods.
The graph-based algorithm [١٠] is one of the first methods developed for feature stream-based online feature selection, using error gradient reduction. The Alpha-Investing algorithm [١١] is an adaptive method that dynamically adjusts the error threshold necessary to accept new features. OSFS [١٧], on the other hand, uses Markov chains and information theory to perform feature selection in datasets with streaming features. Another approach, Online Feature Selection from the Perspective of Uneven Sets (OS-NRRSAR-SA), is based on the fact that data mining with RS (Recommender Systems) requires no additional domain knowledge other than the provided dataset. This method applies classical importance analysis concepts in RS theory to control the unknown feature space in online feature selection problems. It has been evaluated on high-dimensional datasets and shows effectiveness in terms of density, classification accuracy, runtime, and resilience to disruption. This method does not require any extra knowledge and is capable of removing redundant features as they appear [١٨] .
Additionally, OSFSMI and OSFSMI-k algorithms [١٩] make use of mutual information in a streaming fashion to evaluate feature correlation and redundancy in complex classification tasks. These methods do not rely on any learning model during the search process and are classified as filter-based methods. While all of these online feature selection algorithms are designed for single-label data, there is a limited number of methods for online multi-label feature selection, particularly those that optimize multiple criteria during the selection process. In fact, we have not found any methods for multi-label feature selection with streaming samples.
Several individual online multi-label feature selection methods have been proposed, such as MUCO [٢٠], which is based on fuzzy mutual information. The quality of a feature in this method is assessed using fuzzy mutual information, designed to account for label correlation. Another

method, OM-NRS [۲۱], offers an online feature selection approach for multi-label data using an uneven set, proposing a feature subset that includes strong features. This method suggests the nearest neighbor for binning all samples, solving the partial selection problem in uneven regions. A batch version of this algorithm, called FM-NRS, assumes access to the entire data space. Furthermore, MMOFS [۲۲] automatically selects the best feature subset suitable for multi-label classification. The method operates in three phases: the first phase applies a particle swarm optimization technique for a group of input features in a multi-objective framework. The second phase checks for redundancy among selected features compared to previously chosen ones. In the third phase, it identifies and discards features that are irrelevant to selecting new features.

Generally, all the previously introduced methods assume that features are added to the dataset one by one, sequentially. However, in real-world applications, features often have a group structure. In response to this, two methods for online group feature selection are introduced. These methods perform the feature selection process at the group level. Consider $X = [x_1, x_2, \ldots, x_{\mathrm{n}}]^{\mathrm{T}} \in R^{\mathrm{nxd}}$, representing the training dataset with n samples and d-dimensional features, $F = [f_1, f_2, \ldots, f\_d]^{\mathrm{T}} \in R^{\mathrm{d}}$, and the class label $C = [c_1, c_2, \ldots, c\_m]^{\mathrm{T}} \in R^{\mathrm{m}}$. Let $G = \{G_1, G_2, \ldots, G_{\mathrm{n}}\}$ represent non-overlapping groups in the data. The main challenge in these methods is how to optimally select both features within each group and between groups simultaneously. To address this, several feature selection methods have been proposed for group feature streaming, and the details of these methods are as follows.

GFSSF [۲۳] is a method that uses information theory and mutual information to perform well on both group-based and individual training data. It consists of feature-level selection and group-level selection. Initially, it defines concepts like correlation, redundancy, and dependency among features. The algorithm assumes that $I(X;Y)$ represents the mutual information between X and Y and uses this definition in the feature selection process.

OGFS [۲۴] is an efficient feature selection algorithm that utilizes initial group information. It has two main phases: online intra-group selection and online inter-group feature selection. These phases continue until no new features are added.

Group-SAOLA [۸] is an extension of the SAOLA algorithm and is capable of identifying feature groups that are scattered both at the feature and group levels.

These online feature selection algorithms are primarily designed for single-label data. However, there are also well-known algorithms for online feature selection for multi-label data.

OMGFS [25] includes two phases: online group selection and online inter-group selection. These phases continue until no new features are added. In this method, the importance of the feature group is considered during the group selection phase, and redundancy of features is addressed during the inter-group selection phase. However, this method is not suitable when a subset of features within a group is redundant or irrelevant.

MLOSMI [26] starts by clustering the labels. Labels within the same cluster have high correlation, and labels in different clusters are either mutually independent or weakly correlated. Each label cluster is transformed into a multi-class label, reducing the original labels to a lower-dimensional space while considering high-order correlations. Furthermore, feature correlations and redundancy are defined using mutual information to guide the feature selection process. Finally, features are selected online based on the new label space.

These methods provide robust solutions for online feature selection in both single-label and multi-label datasets, taking into account feature group structures and the need for dynamic, scalable, and efficient feature selection processes.

3. Proposed method

In this section, we present the details of the proposed algorithm, named OSM-MI, in which features are gradually added to the dataset one by one over time. Since different input sequences can affect the feature selection algorithm, the features are introduced randomly, and the final results are based on the average of the various sequences provided. Generally, this method includes three main phases, each of which is explained in detail below.

As we know, the goal of feature selection is to choose a compressed subset of features that retains the ability to distinguish the original feature space. Based on information theory, Bell and Wang [۲۷] introduced the first obvious method for selecting a subset.

Principle 1: Given a dataset described by features F and the label vector C, the subset of features S is desirable if $MI(S;C) = MI(F;C)$.

Principle 2: Given a dataset described by features F and the label vector C, S is a set of desired feature subsets if $S \in S$, minimizing the joint entropy $H(S,C)$based on its predictive ability.

Principles 1 and 2 provide an intuitive description of a good feature subset based on information theory and Occam's razor principle. Unlike multi-label learning, the label space in multi-label learning consists of a set

of labels. The approach suitable for multi-label learning is presented as follows:

Principle 3: For the feature space F and the label space L in multi-label learning, the subset of features S is desirable if $MI(S;L) = MI(F;L)$, considering the multi-label data.

Principle 4: For the feature space F and the label space L in multi-label learning, S is a set of desired feature subsets, such that $S \in S$, which minimizes the joint entropy $H(S,L)$ based on its predictive ability in multi-label data.

These principles form the foundation of the OSM-MI algorithm, ensuring that the feature selection process efficiently handles the complexities of multi-label data and retains relevant feature relationships for accurate predictions.

The last two approaches provide a criterion for selecting a multi-label feature subset, meaning the desired subset S should be optimal and have the minimum joint entropy $H(S,L)$. Additionally, we know that a simple way to achieve a desired subset is to comprehensively evaluate feature subsets using these basic methods. However, due to the exponential complexity, this is not feasible even with a moderate number of candidate features. Therefore, some efficient algorithms have been developed to overcome this issue. In this study [۲۸] two criteria, named maximum correlation and minimum redundancy, are introduced. Using these criteria, one can achieve maximum correlation and minimum redundancy for multi-label feature selection. For example, a candidate feature is considered useful if it is highly correlated with all class labels but not redundant with other features selected for all class labels. As we know, the goal of multi-label feature selection is to choose a set of features that have the highest correlation with all class labels. Initially, when no feature has been selected, the algorithm computes the correlation of incoming features with the label set. If a newly added feature is correlated with the labels, it is added to the selected feature set S; otherwise, it is discarded. The correlation value of a feature with the label set $Rel(f_t, L)$ is calculated as follows:

$$maxRel(f_t, L) where\ Rel = \sum_{l_i \epsilon L} MI(f_t; l_i) \quad (1)$$

From equation (1), the following definitions can easily be derived:

Lemma 1: If the newly added feature f_t and any class label $l_i \in L$ are independent, then the mutual information between the newly added feature f_t and the label L will be minimized.

Proof: If $l_i \in L$ and f_t are independent, $MI(l_i|f_t) = 0$. Therefore, $MI(L|f_t) = 0$. Additionally, we have $MI(l_i|f_t) \geq 0$. As a result, the mutual information between L and f_t is minimized.

Lemma 2: If each class label $l_i \in L$ is fully determined by f_t, then the mutual information between the newly added feature f_t and the label L will be maximized.

Proof: If each class label $l_i \in L$ is fully determined by f_t, then $MI(l_i|f_t) = H(l_i)$. From equation (1), it can be concluded that $MI(l_i|f_t) \leq \sum_{i=1}^{m} H(l_i)$. Therefore, the mutual information between the newly added feature f_t and the label L is maximized.

Given Lemmas 1 and 2, equation (1) can be used to select the newly added feature that has the highest correlation with all class labels.

A newly added feature, based on its maximum correlation, might cause redundancy. For example, a new feature may be correlated with some previously selected features. On the other hand, we know that if two features are highly dependent, the classification quality will not be significantly affected by removing one of them. Therefore, redundancy between features must be measured during the feature selection process. Unlike traditional single-label feature selection, multi-label feature selection not only includes redundancy between individual features but also considers the pairwise relationship between features for each class label. If S_t is a subset of selected features, the minimum redundancy is defined as follows (Equation 2). In this equation, the first term $\sum_{f_j \in S} MI(f_t; f_j)$ represents the redundancy between the newly added feature f_t and the features selected in S_{t-1}. The second term $\sum_{f_j \in S} \sum_{l_i \in L} MI(f_t; l_i | f_j)$ represents the relationship between the newly added feature f_t and all class labels L, accounting for the conditional redundancy. Combining these two terms shows the conditional redundancy between the candidate feature f_t and the selected features in S_{t-1}.

$$minRed(f_t, S_{t-1}, L)\ where\ Red = \frac{1}{|S_{t-1}|} \sum_{f_j \in S_{t-1}} \left[MI(f_t; f_j) - \sum_{l_i \in L} MI(f_t; l_i | f_j) \right] \quad (2)$$

Combining Maximum Correlation and Minimum Redundancy (MDMR): In this phase, an operator is defined to combine D (correlation) and R (redundancy) and optimize both parameters simultaneously.

$$max\ \delta(Rel, Red), \delta = Rel - Red \quad (3)$$

Based on Equation (3), the importance of feature f_t can be calculated as follows:

$$max \left[\sum_{l_i \epsilon L} MI(f_t; l_i) - \frac{1}{|S_{t-1}|} \sum_{f_j \in S_{t-1}} (MI(f_t; f_j) - \sum_{l_i \in L} MI(f_t; l_i | f_j)) \right] \quad (4)$$

From Equation (4), it can be deduced that the selected feature f_t must maximize $\delta(Rel, Red)$. Moreover, in Equation (4), the term $MI(f_t; l_i)$ is constant for f_t, so the equation simplifies as follows:

(5)

$$
\begin{aligned}
(Rel, Rel) = & \sum_{l_i \epsilon L} MI(f_t; l_i) \\
& - \frac{1}{|S_{t-1}|} \sum_{f_j \in S_{t-1}} \left[MI(f_t; f_j) \right. \\
& \left. - \sum_{l_i \epsilon L} MI(f_t; l_i | f_j) \right] \\
\propto & \ |S| \sum_{l_i \epsilon L} MI(f_t; l_i) - \sum_{f_j \epsilon S} MI(f_t; f_j) \\
& - \sum_{f_j \epsilon S} \sum_{l_i \epsilon L} MI(f_t; l_i | f_j) \\
= & \sum_{f_j \epsilon S} \left[\sum_{l_i \epsilon L} MI(f_t; l_i) - MI(f_t; f_j) \right. \\
& \left. + \sum_{l_i \epsilon L} MI(f_t; l_i | f_j) \right] \\
= & \sum_{f_j \epsilon S} \left[\sum_{l_i \epsilon L} MI(f_t; l_i) - MI(f_t; f_j) \right. \\
& \left. + \sum_{l_i \epsilon L} MI(f_t; l_{i(i=1,2,\ldots,|L|)} | f_j) \right] \\
= & \sum_{f_j \epsilon S} \left[\sum_{l_i \epsilon L} MI(f_t; l_i) \right. \\
& \left. - MI(f_i, l_{i(i=1,2,\ldots,|L|}, f_j) \right] \\
= & \sum_{f_j \epsilon S} \sum_{l_i \epsilon L} [MI(f_t; l_i) - MI(f_i, l_i, f_j)].
\end{aligned}
$$

From Equation (5), we can see that the first term focuses on the correlation between the candidate feature and all class labels, while the second term specifies the conditional redundancy between the candidate feature and the selected features. Therefore, the MDMR criterion can be used to rank a set of features and determine the best newly added feature f_t . The newly added feature must have the highest value of the difference between Rel and Red. In other words, when a new feature f_t is introduced, it gains a "fitness" value based on its correlation with the labels (Rel) and redundancy (Red) with previously selected features. If the number of selected features equals the size previously specified by the user, one of the features will be removed, and the new feature will replace it. Thus, the newly added feature is compared with all previously selected features $g \in S_t$, and if a feature has a lower fitness value than the newly added feature, it is replaced.

Figure (1) shows the pseudocode of the proposed method.

Input: f_t is the newly arrival feature f at time t. λ is the fitness function, S_0:{ }, k: Size of selected of features.

Output: The selected feature subset till time t.

Begin algorithm

$f_t \leftarrow$ newly arrival feature at time t.

// Checking for dependency of new arrival feature f_t.

Compute Rel_{f_t} .

// Checking for redundancy features in S_t.

Compute Red_{f_t} .

// Checking for fitness features in S_t.

Compute $fitness_{f_t S_t}$.

$max = fitness_{f_t S_t}$

$N = f_t$

If Size (S_t) >= k

For each feature $g \in S_t$

If (max > $fitness_{g S_t}$) then

$g \leftarrow N$ and remove feature g .

Else If (max < $fitness_{g S_t}$) then

remove feature N .

End if

End if

End for

Else f_t add to S_t.

Update $fitness_{g S_t}$ for each feature $g \in S_t$.

Until no new feature are available.

Return S_t;

Figure 1. Pseudocode of the proposed method

4. Analysis and Experiments

This section presents the results of ten different feature entry sequences across all datasets. In all these tables, the columns represent online feature selection algorithms, and the rows correspond to a dataset. The best value in each row is highlighted in bold and underlined. The last row shows the statistical results obtained from the Wilcoxon test. The Wilcoxon test is used to compare the performance of feature selection methods. It is an inferential statistical test used to assess the similarity between two related samples with a rank scale. This test calculates the p-value for each data pair

and analyzes the differences. In comparing feature selection methods, the null hypothesis indicates that there is no difference in the performance of the two feature selection methods. If the p-value is less than or equal to a specified significance level ($\alpha = 0.05$), the null hypothesis is rejected, and it can be concluded that there is a significant difference between the two methods [٢٩]. One column of each table presents the statistical comparison of the proposed method with other methods. A positive sign indicates that the proposed method outperforms the other feature selection methods, while a negative sign indicates that the proposed method is not superior, and the (=) sign indicates that there is no significant difference between the performance of the two feature selection methods.

Tables (1-6) show the accuracy, hamming loss, one-error, coverage, average precision, and rank loss obtained using the ML-kNN classifier. From the results of these tables, it can be seen that the proposed algorithm achieves the best accuracy among the other methods.

Table 1. Comparison of the accuracy of the OSM-MI method with other multi-label streaming feature selection methods.

	OSM-MI	OM-NRS	OMGFS	MUCO
Yeast	**0/5698**	0/5112	0/5214	0/5024
Medical	**0/5334**	0/5325	0/5521	0/5145
Scene	**0/5301**	0/5021	0/5298	0/4954
Enron	**0/3630**	0/3218	0/3512	0/3008
Genbase	**0/9078**	0/9010	0/9024	0/8825
Image	**0/4176**	0/3458	0/4154	0/3947
Bibtex	**0/1307**	0/1012	0/1287	0/1102
Corel5k	**0/1907**	0/1662	0/1886	0/1784
Wilcoxon		+	+	+

Table 2. Comparison of the Hamming-loss of the OSM-MI method with other multi-label streaming feature selection methods.

	OSM-MI	OM-NRS	OMGFS	MUCO
Yeast	**0/1978**	0/1995	0/2084	0/2101
Medical	**0/0174**	0/0201	0/0188	0/0195
Scene	0/1307	0/1543	0/1452	**0/1301**
Enron	0/0514	0/0521	**0/0512**	0/063
Genbase	**0/0049**	0/0112	0/0058	0/0107
Image	**0/6100**	0/9254	0/6301	0/8839
Bibtex	0/0104	0/0140	**0/0102**	0/0152
Corel5k	**0/0095**	0/0095	0/0098	0/0109
Wilcoxon		+	+	+

Table 3. Comparison of the One-error of the OSM-MI method with other multi-label streaming feature selection methods.

	OSM-MI	OM-NRS	OMGFS	MUCO
Yeast	**0/1998**	0/2431	0/2007	0/2527
Medical	**0/2604**	0/3228	0/2698	0/2978
Scene	**0/3726**	0/5873	0/4125	0/4456
Enron	**0/3267**	0/3455	0/3385	0/3715
Genbase	**0/0106**	0/0352	0/0220	0/0251
Image	0/3671	0/4450	**0/3546**	0/4127
Bibtex	**0/5776**	0/6613	0/6157	0/6309
Corel5k	**0/6887**	0/7535	0/7264	0/7001
Wilcoxon		+	+	+

Table 4. Comparison of the Coverage of the OSM-MI method with other multi-label streaming feature selection methods.

	OSM-MI	OM-NRS	OMGFS	MUCO
Yeast	6/9853	6/6235	**6/4183**	6/6057
Medical	**4/0005**	5/9254	4/1524	5/3289
Scene	**0/7914**	1/6213	0/8503	1/5829
Enron	**14/092**	14/9157	15/1002	14/5780
Genbase	**0/7230**	0/8951	0/7568	0/83259
Image	**0/7586**	1/8997	0/8038	1/2036
Bibtex	**53/2742**	63/5462	55/6322	60/4378
Corel5k	**118/4516**	121/5258	120/3048	119/2171
Wilcoxon		+	+	+

Table 5. Comparison of the Precision of the OSM-MI method with other multi-label streaming feature selection methods.

	OSM-MI	OM-NRS	OMGFS	MUCO
Yeast	**0/7620**	0/7041	0/7545	0/7350
Medical	**0/7ˇ64**	0/6522	0/7452	0/7616
Scene	**0/7861**	0/7053	0/7540	0/7750
Enron	**0/6^87**	0/6349	0/6336	0/6450
Genbase	**0/9^54**	0/9673	0/9185	0/9720
Image	**0/7571**	0/7313	0/7502	0/7446
Bibtex	0/3958	0/3224	**0/4002**	0/3854
Corel5k	**0/2494**	0/2214	0/2420	0/2340
Wilcoxon		+	+	+

Table 6. Comparison of the Ranking loss of the OSM-MI method with other multi-label streaming feature selection methods.

	OSM-MI	OM-NRS	OMGFS	MUCO
Yeast	**0/1783**	0/2008	0/1832	0/2014
Medical	0/0436	0/1184	0/0910	**0/0107**
Scene	**/0/2228**	0/3562	0/2366	0/3098
Enron	**0/0958**	0/1103	0/0937	0/1003
Genbase	0/0110	0/0615	**0/0106**	0/0125
Image	**0/1950**	0/3164	0/1997	0/2389
Bibtex	**0/2134**	0/2904	0/2256	0/2507
Corel5k	**0/1359**	0/1493	0/1456	0/1415
Wilcoxon		+	+	+

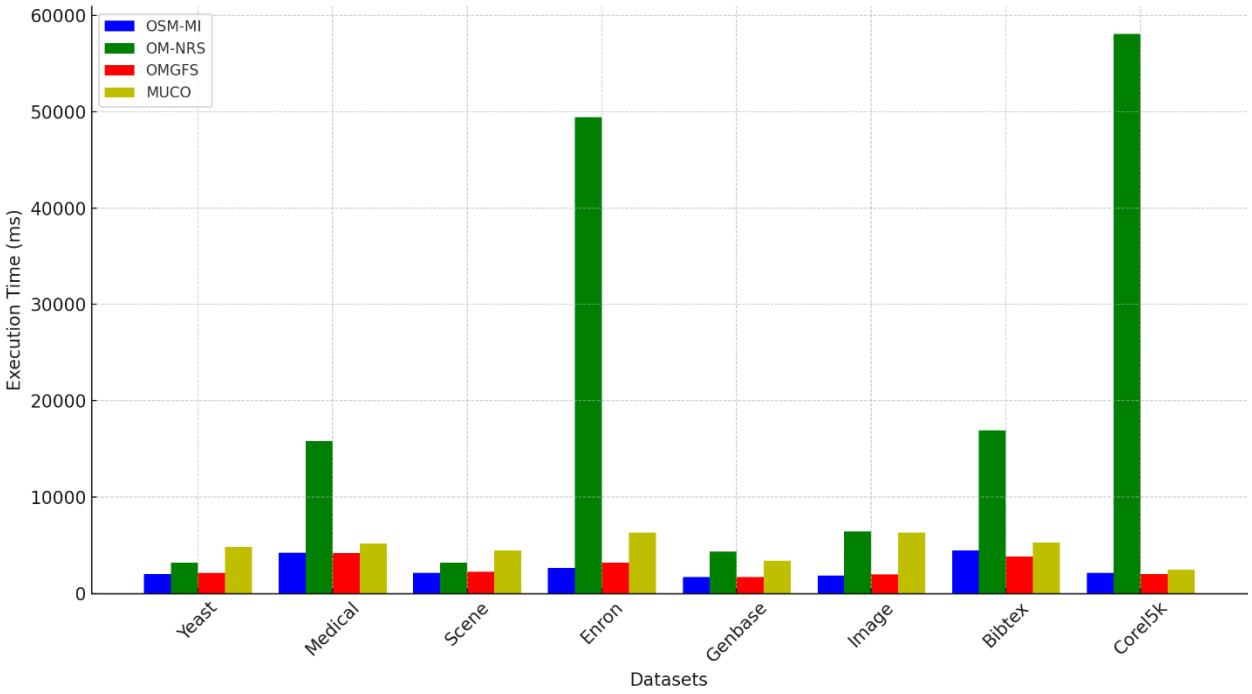

Figure 1. Comparison of Execution Times for OSM-MI and Other Multi-label Streaming Feature Selection Methods

In terms of accuracy, OSM-MI consistently outperforms the other methods in most datasets. For example, in the Yeast dataset, OSM-MI achieves an accuracy of 0.5698, which is higher than the others. Similarly, in the Enron dataset, OSM-MI's accuracy is 0.3630, significantly better than the other methods, demonstrating better generalization and performance in multi-label streaming feature selection tasks.

Regarding Hamming loss, OSM-MI shows superior performance by achieving lower values across most datasets. In the Yeast dataset, the Hamming loss of OSM-MI is 0.1978, lower than that of the other methods, indicating that OSM-MI is better at minimizing incorrect labels. The Genbase dataset shows the lowest Hamming loss for OSM-MI at 0.0049, further supporting its effectiveness in multi-label classification. For the one-error metric, which measures the fraction of times the top-ranked label is incorrect, OSM-MI again outperforms the other methods. In the Genbase dataset, OSM-MI achieves a one-error of 0.0106, significantly outperforming the alternatives. This demonstrates that OSM-MI is better at minimizing incorrect top predictions, which is crucial in multi-label tasks where the correct top label is prioritized.

In terms of coverage, which measures the fraction of relevant labels ranked in the top positions, OSM-MI also performs well. It maintains high coverage values, with Yeast achieving 9853/6 and Corel5k 118/4516. This indicates that OSM-MI is effective at ensuring a larger proportion of relevant labels are included in the top positions compared to other methods.

Regarding precision, OSM-MI shows competitive results. In the Yeast dataset, OSM-MI achieves a precision of 0.7620, outperforming the other methods. This suggests that OSM-MI is effective at making accurate predictions, particularly in terms of the proportion of correct labels.

Finally, in terms of ranking loss, OSM-MI demonstrates strong performance in minimizing the ranking loss. For example, in the Yeast dataset, it achieves a ranking loss of 0.1783, better than the other methods. This indicates that OSM-MI effectively ranks the relevant labels higher, which is essential in multi-label tasks where the order of predictions matters.

The Wilcoxon test results indicate that OSM-MI outperforms the other methods in several datasets, as shown by the "+" sign in the Wilcoxon row. This statistical significance further supports the effectiveness of OSM-MI in multi-label streaming feature selection tasks.

In conclusion, the OSM-MI method consistently outperforms OM-NRS, OMGFS, and MUCO across multiple evaluation metrics, including accuracy, hamming loss, one-error, ranking loss, precision, and coverage. These results highlight the robustness and efficiency of OSM-MI as a method for multi-label streaming feature selection, demonstrating its superiority in a variety of datasets. The statistical significance of the results, supported by the Wilcoxon test, underscores the effectiveness of the OSM-MI approach.

Figure 2 presents a comparison of the execution times of the OSM-MI method with other multi-label streaming feature selection methods, including OM-NRS, OMGFS, and MUCO, across various datasets. The datasets used in the comparison include Yeast, Medical, Scene, Enron, Genbase, Image, Bibtex, and Corel5k. Each method's execution time is represented by a distinct colored bar, with OSM-MI shown in blue, OM-NRS in green, OMGFS in red, and MUCO in yellow.

As observed, OSM-MI consistently demonstrates lower execution times compared to the other methods across most datasets. For example, in the Yeast dataset, OSM-MI achieves an execution time of 2023 ms, significantly outperforming the other methods, such as MUCO, which has a higher execution time of 4849 ms. This trend is consistent across other datasets, where OSM-MI generally exhibits faster execution times, suggesting its efficiency in handling multi-label feature selection tasks. In some cases, such as the Enron dataset, the difference in execution times is substantial, with OSM-MI performing much better than OM-NRS and MUCO. The figure highlights the overall efficiency of the OSM-MI method in terms of execution time, making it a preferred choice for large-scale multi-label streaming feature selection tasks. The consistency of OSM-MI's performance across various datasets reinforces its robustness and suitability for real-time applications.

5. Conclusion and Future Work

In this study, we proposed the OSM-MI method for multi-label streaming feature selection and evaluated its performance against other well-established methods, including OM-NRS, OMGFS, and MUCO. The results showed that OSM-MI outperforms the other methods in terms of accuracy, hamming loss, one-error, precision, and ranking loss across a variety of datasets. Additionally, OSM-MI demonstrates superior execution times, making it an efficient choice for real-time applications. The statistical significance of these results, supported by the Wilcoxon test, further confirms the effectiveness of the OSM-MI method in multi-label streaming feature selection tasks. The proposed method not only ensures high classification accuracy but also maintains low redundancy and maximizes feature relevance in multi-label data.

Future work could focus on improving scalability with parallel computing, exploring deep learning techniques, and enhancing robustness to noisy data. Additionally, testing the method in other multi-label tasks, like image or text classification, would help assess its versatility.

References

[1] S. Gilpin, B. Qian, and I. Davidson, “Efficient hierarchical clustering of large high dimensional datasets,” in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, San Francisco, California, USA, 2013, pp. 1371-1380. https://doi.org/10.1145/2505515.2505527

[2] J. Dai, W. Chen, and Y. Qian, “Multi-label feature selection with missing features via implicit label replenishment and positive correlation feature recovery,” *IEEE Transactions on Knowledge and Data Engineering*, 2025. 10.1109/TKDE.2025.3536080

[3] A. RAFIEI, P. MORADI, and A. Ghaderzadeh, “Multi-Label Feature Selection Using a Hybrid Approach Based on the Particle Swarm Optimization Algorithm,” 2023. 20.1001.1.16823745.1401.20.4.7.7

[4] P. Kiyoumarsi, F. Kiyoumarsi, B. Z. Dehkordi, and M. Karbasiyoun, “A Feature Selection Method on Gene Expression Microarray Data for Cancer Classification Abstract,” *Journal of Optimization in Soft Computing*, vol. 2, no. 3, pp. 35-44, 2024. https://doi.org/10.82553/josc.2024.140308101189068

[5] J. Abdollahi, B. Nouri-Moghaddam, N. Mikaeilvand, S. J. Gudakahriz, A.

Khosravani, and A. Mirzaei, "A Review of Feature Selection," *Journal of Optimization in Soft Computing*, vol. 2, no. 4, pp. 16-20, 2025. https://doi.org/10.82553/josc.2025.140309071191740

[6] W. Ding, T. F. Stepinski, Y. Mu, L. Bandeira, R. Ricardo, Y. Wu, Z. Lu, T. Cao, and X. Wu, "Subkilometer crater discovery with boosting and transfer learning %J ACM Trans. Intell. Syst. Technol," vol. 2, no. 4, pp. 1-22, 2011. https://doi.org/10.1145/1989734.1989743

[7] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal Graph-Based Reranking for Web Image Search %J Trans. Img. Proc," vol. 21, no. 11, pp. 4649-4661, 2012. 10.1109/TIP.2012.2207397.

[8] K. Yu, X. Wu, W. Ding, and J. Pei, "Scalable and Accurate Online Feature Selection for Big Data %J ACM Trans. Knowl. Discov. Data," vol. 11, no. 2, pp. 1-39, 2016. https://doi.org/10.1145/2976744

[9] Y. Hochma, and M. Last, "Fast online feature selection in streaming data," Machine Learning, vol. 114, no. 1, pp. 1, 2025. https://doi.org/10.1007/s10994-024-06712-x

[10] S. Perkins, and J. Theiler, "Online feature selection using grafting," in Proceedings of the Twentieth International Conference on International Conference on Machine Learning, Washington, DC, USA, 2003, pp. 592-599.

[11] J. Zhou, D. P. Foster, R. A. Stine, and L. H. Unga, "Streamwise feature selection,".Journal of Machine Learning Research, vol. 7, pp. 1861-1885, 2006.

[12] L. Zou, T. Zhou, and J. Dai, "Online Multi-Label Streaming Feature Selection by Label Enhancement and Fuzzy Synergistic Discrimination Information," IEEE Transactions on Fuzzy Systems, 2025. 10.1109/TFUZZ.2025.3554982

[13] J. Liu, Y. Lin, Y. Li, W. Weng, and S. Wu, "Online Multi-label Streaming Feature Selection Based on Neighborhood Rough Set," vol. 84, pp. 273-287, 2018. https://doi.org/10.1016/j.patcog.2018.07.021

[14] J. Liu, Y. Lin, S. Wu, and C. Wang, "Online Multi-label Group Feature Selection," Knowledge-Based Systems, vol. 143, pp. 42-57, 2018. https://doi.org/10.1016/j.knosys.2017.12.008

[15] W. Jiang, G. Er, and Q. Dai, "Similarity-based online feature selection in content-based image retrieval," in IEEE TRANSACTIONS ON IMAGE PROCESSING, 2006, pp. 02–712. 10.1109/TIP.2005.863105

[16] A. Rafie, P. Moradi, and A. Ghaderzadeh, "A multi-objective online streaming multi-label feature selection using mutual information," Expert Systems with Applications, vol. 216, pp. 119428, 2023. https://doi.org/10.1016/j.eswa.2022.119428

[17] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online feature selection with streaming features," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, pp. 1178–1192. 10.1109/TPAMI.2012.197

[18] S. Eskandari, and M. M. Javidi, "Online streaming feature selection using rough sets," International Journal of Approximate Reasoning, vol. 69, pp. 35-57, 2016. https://doi.org/10.1016/j.ijar.2015.11.006

[19] M. Rahmaninia, and P. Moradi, "OSFSMI: online stream feature selection method based on mutual information," Applied Soft Computing, vol. 68, pp. 733-746, 2018. https://doi.org/10.1016/j.asoc.2017.08.034

[20] Y. Lin, Q. Hu, J. Liu, J. Li, and X. Wu, "Streaming feature selection for multilabel learning based on fuzzy mutual information," IEEE Transactions on Fuzzy Systems, vol. 25, no. 6, pp. 1491-1507, 2017. 10.1109/TFUZZ.2017.2735947

[21] J. Liu, Y. Lin, Y. Li, W. Weng, and S. Wu, "Online multi-label streaming feature selection based on neighborhood rough set," Pattern Recognition, vol. 84, pp. 273-287, 2018. https://doi.org/10.1016/j.patcog.2018.07.021

[22] D. Paul, A. Jain, S. Saha, and J. Mathew, "Multi-objective PSO based online feature selection for multi-label classification," Knowledge-Based Systems, vol. 222, pp. 106966, 2021. https://doi.org/10.1016/j.knosys.2021.106966

[23] H. L. X. W. Z. L. W. Ding, "Group Feature Selection with Streaming Features," in 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA 2013. 10.1109/ICDM.2013.137

[24] J. Wang, M. Wang, P. Li, and L. Liu, “Online Feature Selection with Group Structure Analysis,” EEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, 2015. https://doi.org/10.48550/arXiv.1608.05889

[25] X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” Advances in neural information processing systems, vol. 18, 2005.

[26] H. Wang, D. Yu, Y. Li, Z. Li, and G. Wang, "Multi-label online streaming feature selection based on spectral granulation and mutual information." pp. 215-228. https://doi.org/10.3390/e25071071.

[27] S. C. H. Hoi, J. Wang, P. Zhao, and R. Jin, “Online feature selection for mining big data,” in Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Beijing, China, 2012, pp. 93-100. https://doi.org/10.1016/j.swevo.2025.101896

[28] L. Yu, and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution." pp. 856-863.

[29] H. O. Parametric, “Handbook Of Parametric And Nonparametric Statistical Procedures.” https://doi.org/10.1201/9780429186196.