

I-mRMR: Incremental Max-Relevance, and Min-Redundancy Feature Selection*

ABSTRACT

An incremental method of feature selection based on mutual information, called incremental Max-Relevance, and Min-Redundancy (I-mRMR), is presented. I-mRMR is an incremental version of Max-Relevance, and Min-Redundancy feature selection (mRMR), which is used to handle streaming data or large-scale data. First, Incremental Key Instance Set is proposed which composes of the non-distinguished instances by the historical selected features. Second, an incremental feature selection algorithm is designed in which the incremental key instance set, replacing of all the seen instances so far, is used in the process of adding representative features. Since the Key Instance Set is far less than the whole instances, the incremental feature selection by using this key set avoid redundant computation and save computation time and space. Finally, the experimental results show that I-mRMR could significantly or even dramatically reduce the time of feature selection with an acceptable classification accuracy. The main advantage of I-mRMR is that it makes full use of the historical information, reduce the training scale greatly, and save training time.

KEYWORDS

Feature selection, Incremental algorithm, normalized mutual information, min-redundancy, max-relevance.

ACM Reference Format:

. 2019. I-mRMR: Incremental Max-Relevance, and Min-Redundancy Feature Selection. In *Proceedings of ACM International conference on Web Search and Data Mining (WSDM'19)*, Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In recent years, we encounter databases in which the issue of data is too big to loaded in the memory or the data are streaming collected over time. Storing and processing all data might be computationally costly and impractical. To deal with this issue, incremental algorithms have become feasible and effective tool in machine learning and data mining techniques [3, 4, 18, 23, 26].

Incremental learning is a promising approach to refreshing data mining results, which utilizes previously saved results or data structures to avoid the expense of re-computation [10, 14, 30, 33, 34]. Incremental feature selection (sometimes also called online feature selection) is one important part of “incremental machine learning”

fields [12, 16]. The main idea of incremental feature selection is that only part of the data are to be considered at one time and the results are subsequently combined. Thus incremental feature selection technique makes full use of the historical information, reduce the training scale greatly, and save training time [29].

Feature selection based on mutual information, as a known way of feature selection, have been deeply studied [5, 6, 8, 11, 17, 20, 28], since mutual information (MI) [2, 19, 24] is a good tool to measure the correlation and redundancy among features. As a pioneer, Battiti [1] proposed a greedy selection method called MIFS based on mutual information between inputs and outputs. Considering MIFS does not work well in nonlinear problems, Kwak and Choi [20] proposed an improved feature selection method MIFS-U which is feasible and effective on nonlinear applications. However, both Battiti and Kwak's methods omit the redundancy among features, only relevance among features and labels are considered. Peng et al. [22] then proposed a heuristic “Max-Relevance and Min-Redundancy” framework for feature selection. In [22] it is pointed that mRMR criterion is equal to max-dependency. Furthermore, Pablo and Tesmer [9] proposed an updated feature selection method, called normalized mutual information features selection. This method updated the Max-Relevance and Min-Redundancy criterion and numerical experiments demonstrates that it is faster and better than MIFS, MIFS-U. However, there exist one common limitation among the above mentioned methods. That is, most of them could only be applied to static data. When new instances are arriving successively, these methods have to be re-computed on the updated datasets.

To select features on streaming datasets, some incremental feature selection algorithms have been proposed based on entropy or its generalization [15, 21, 27]. Whereas max-relevance and min-redundancy based MI is not considered yet. Now, it is promising to design an incremental feature selection method based on max-relevance and min-redundancy.

In this paper, we propose an incremental feature selection algorithm, called Incremental Max-Relevance, and Min-Redundancy Feature Selection (I-mRMR). First, Incremental Key Instance Set is proposed which is composed of part of instances not distinguished by historical selected features. An incremental algorithm is then proposed based on this Incremental Key Instance Set. Finally, the main advantage of I-mRMR is that it makes full use of the historical selected features, reduce the training scale greatly, and save training time.

The remainders of this paper are organized as follows. Section 2 reviews MI and mRMR based on normalized mutual information. Section 3 introduces the concept of Incremental Key Instance Set and presents the incremental feature selection algorithm I-mRMR. In Section 4, ten UCI Datasets and two extremely high dimensional datasets are employed to illustrate the effectiveness and efficiency of I-mRMR. Section 5 concludes this paper.

*Produces the permission block, and copyright information

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WSDM'19, February 2019, Melbourne Australia

© 2016 Copyright held by the owner/author(s).

ACM ISBN 123-4567-24-567/08/06.

https://doi.org/10.475/123_4

2 PRELIMINARIES

In this section, MI and mRMR are reviewed. For more detailed information about them, please kindly refer to [9, 22, 25].

2.1 notation description

Given a set of original instances $U = [x^{(1)}, x^{(2)}, \dots, x^{(n)}]^T$. Here $U \in R^{(n \times p)}$ is a matrix with n is the number of original instances and p is the number of all features. $x^{(i)} \in R^p$ is a row vector representing the i -th instance in U . S is the index set of selected feature subset. \bar{S} denotes the complementary set of S . x_t is a column vector representing the t -th feature. $x_S^{(i)}$ represents a vector of $x^{(i)}$ under feature subset $S(i = 1, \dots, n)$. $Y = [y^{(1)}, \dots, y^{(n)}]^T$ is a column vector representing the label feature in U . Here $y^{(i)}$ is the label for the i -th instance in $U(i = 1, \dots, n)$.

2.2 Mutual information

MI is often used to measure the relevance of two random variables, which is defined as follows.

Formally, given two discrete random variables X and Y , MI between X and Y with a joint probability mass function $p(x, y)$ and marginal probabilities $p(x)$ and $p(y)$ is defined as follows:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} \log \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

The larger the value of MI, the larger the relevance of X and Y is. Considering features as variables, the larger the value of MI, the larger the relevance among features is, the more instances with different labels could be distinguished.

Just as described in [25], MI can be written as the following equation:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (2)$$

Where $H(X) = -\sum p(x) \log p(x)$ is the information entropy, $H(X|Y) = -\sum \sum p(x, y) \log p(x|y)$ is the conditional entropy. It is easy to derive the range of mutual information $I(X; Y)$:

$$0 \leq I(X, Y) \leq \min\{H(X), H(Y)\}$$

MI is often used to describe the relevance degree between features and label [1, 2, 19, 24]. It is preferential to select the features with the maximum MI to add to the feature subset candidate. Since every feature could be seen as a variable, we use the notation of feature to replace the notion of variables when no confusion arises in the following of this paper.

2.3 Max-Relevance and Min-Redundancy

Max-Relevance is to find the feature x_t that satisfies the following formula:

$$\max_{t \in S} D(S), \text{ where } D = \frac{1}{|S|} \sum_{t \in S} I(Y; x_t) \quad (3)$$

By the Max-Relevance criterion, only the relevance between the features and labels are considered, whereas the relevance among the features is not considered. Thus there may exist great redundancy among the selected features. As a result, it is necessary to

make the redundancy among the selected features as small as possible.

$$\min_{t \in S} R(S), \text{ where } R = \frac{1}{|S|^2} \sum_{k, t \in S} I(x_k, x_t) \quad (4)$$

The above two criteria are combined, called "Max-Relevance and Min-Redundancy", and defined as follows.

$$\max \Phi(D, R), \Phi = D - R \quad (5)$$

Suppose that the feature subset candidate we have selected so far is S_{m-1} , and $m - 1$ indicates that $m - 1$ features have been selected. And then the feature with the maximum value of $\Phi(D, R)$ is selected. The incremental feature selection algorithm optimizes the following formula:

$$\max_{k \in F - S_{m-1}} [I(Y, x_k) - \frac{1}{|S_{m-1}|} \sum_{t \in S_{m-1}} I(x_k, x_t)] \quad (6)$$

Based on this idea, a series of feature selection methods based on max-relevance and min-redundancy have been done, please kindly refer to [7, 13, 22, 32]. Among these works, normalized mutual information based mRMR is the state-of-the-art work, and then in the following we briefly review it, based on which, we design an incremental feature selection method.

2.4 Normalized mutual information feature selection

Since the criteria of the relevance between x_k and x_t are not uniform, the normalized mutual information is then used as the evaluation criterion to measure the degree of feature relevance.

The normalized mutual information $NI(x_k, x_t)$ between the feature x_k and the feature x_t is then defined as follows.

$$NI(x_k, x_t) = \frac{I(x_k, x_t)}{\min\{H(x_k), H(x_t)\}} \quad (7)$$

Therefore, "Max-Relevance and Min-Redundancy" criterion can be rewritten as follows:

$$\max_k [I(Y, x_k) - \frac{1}{|S_{m-1}|} \sum_{t \in S_{m-1}} NI(x_k, x_t)] \quad (8)$$

The algorithm based on normalized mutual information feature selection is formulated in Algorithm 1. In this paper, we briefly notate this algorithm mRMR when no confusion arises.

3 THE PROPOSED INCREMENTAL ALGORITHM

When some new instances arriving, the classical feature selection algorithm has to be recomputed on all the seen data so far, which is really time and computation consuming. To solve this problem, we propose an incremental algorithm based on the "Max-Relevance and Min-Redundancy" criterion. The key idea of our proposed method is to update and maintain the previously selected feature subset by finding the features more representative for discriminating the new instances from its current surrounding.

Algorithm 1: Normalized Mutual Information Feature Selection(mRMR)Input: U , the given number of selected features K , $F=1, \dots, p$, $S=\emptyset$ Output: the set S containing the selected featuresStep 1: Compute $I(Y; x_k)$ for every $k \in F$.Step 2: Select the first feature $k^* = \arg_k \max I(Y; x_k)$.Step 3: Set $S \leftarrow S \cup \{k^*\}$, $F \leftarrow F - \{k^*\}$.Step 4: While $|S| < K$ do.

{

 Compute $I(x_k; x_t)$ for all pairs $(x_k; x_t)$, with $k \in F$ and $t \in S$

 if it is not available;

 Select feature $k^* \in F$ that maximizes measures (8);

 Set $S \leftarrow S \cup \{k^*\}$, $F \leftarrow F - \{k^*\}$.

}

Step 5: Return S .

3.1 Problem Definition

When some new instances, represented by $\Delta U \in R^{m \times p}$ (where m represents the number of newly added instances), are added to U , $y^{(n+j)}$ is the label for the j -th instance in ΔU , $j = 1, \dots, m$. The selected feature subset S has to be updated from U to $U \cup \Delta U$. The traditional method is directly to recompute the feature selection method on all seen instances $U \cup \Delta U$ to obtain the updated feature subset $S_{U \cup \Delta U}$. It is very time and space consuming and many redundant computations are conducted. Therefore, it is necessary to reduce the amount of computation by using some incremental mechanisms. And we refer to feature selection method based incremental mechanism as incremental feature selection.

An incremental feature selection algorithm is then proposed based on normalized mutual information. The purpose of this algorithm is to dynamically find the selected feature subset $S_{U \cup \Delta U}$.

3.2 Incremental Key Instance Set

To incrementally update the selected feature subset S , it is necessary to find the features more representative for discriminating the new instances from its current surrounding. That is to say, it is fundamental to find those instances which are key to select the new representative features.

In the following we propose a concept called Incremental Key Instance Set which composes of part of the seen instances so far which are undistinguished by the original features subset S .

DEFINITION 1. Given U , S , and ΔU , then Incremental Key Instance Set of S , denoted by ΔI_S , is defined as follows.

$$\begin{aligned} \Delta I_S = & \{x^{(i)} \in U | \exists x^{(n+j)} \text{ s.t. } x_S^{(i)} = x_S^{(n+j)}, y^{(i)} \neq y^{(n+j)}\} \cup \\ & \{x^{(n+j)} \in \Delta U | \exists x^{(i)} \in U \text{ s.t. } x_S^{(i)} = x_S^{(n+j)}, y^{(i)} \neq y^{(n+j)}\} \end{aligned} \quad (9)$$

Incremental Key Instance Set ΔI_S composes of such instances which have the same feature values on S but the different labels, which means that the features in S could not distinguish the new instances from its current surrounding and then some new features should be added. ΔI_S plays the key role to find the new features.

A function that measures the significance of the feature according to the criterion of the "Max-Relevance and Min-Redundancy"

is then proposed based on Increment Key Instance Set. Since Increment Key Instance Set is far less than all the seen instances so far, it greatly reducing the computation, time and space consumption when updating feature selection.

DEFINITION 2. Given U , Y , F and S , for every $k \in \bar{S}$ and $t \in S$, the significance degree of x_k with respect to Y and S is defined as follows.

$$Sig(x_k, S, Y) = I(Y, x_k) - \frac{1}{|S|} \sum_{t \in S} NI(x_k, x_t) \quad (10)$$

Computing the significance degrees of \bar{S} on ΔI_S , all the features in \bar{S} are then sorted. Thus the feature with the maximum distinguishing power, i.e. maximum significance degree, is added to S .

3.3 Stop criterion

Many of the previous stop criterion for MI based feature selection algorithms are to select top K . One main advantage of this approach is that it is easy to determine the number of selected features. But its shortcoming is obvious, that is, there is no enough reason to believe that the selected K features could completely distinguish all the instances. In this paper we would like to propose a stop criterion which avoids this shortcoming.

Usually, the conditional entropy $H(Y|S)$ is used to measure the significance of S . Here the feature subset S functions as a random variable. $H(Y|S) = 0$ means that the selected feature subset S could definitely distinguish all the instances so far. The larger $H(Y|S)$ is, some more instances could not be distinguished by S .

The mutual information of the selected feature subset S is $I(Y; S) = H(Y) - H(Y|S)$. The relationship between mutual information and conditional entropy is as follows.

If $H(Y|S) = 0$, then $I(Y; S) = H(Y)$

Once $I(Y; S)$ is equal to or very close to the information entropy of Y , the selected feature subset S could almost distinguish all the instances.

The stopping criterion of I-mRMR is then set as follows:

$$I(Y; S) = H(Y) \quad (11)$$

3.4 Incremental feature selection algorithm

In this subsection, we present the incremental feature selection algorithm when a set of new instances arriving. I-mRMR is designed in algorithm 2.

3.5 Scalability Analysis

The time complexity of the algorithm is shown in Table 1. The main step of I-mRMR is to find the Incremental Key Instance Set ΔI_S , which gradually decreases the computation of choosing new representative features.

From Table 1, it is to see that When the dimension of the datasets is high, the time complexity of I-mRMR is $O(|\Delta I_S|(\log |\Delta I_S|)|F - S|)$. Comparing with the time complexity of mRMR, I-mRMR spends less time because $|\Delta I_S|$ is far smaller than $|U + \Delta U|$. In the following numerical experimental part, we experimentally demonstrate the efficiency of I-mRMR.

Algorithm 2: An incremental algorithm for feature selection based on Max-Relevance, and Min-Redundancy (I-mRMR)

Input: $U, F, Y, S, \Delta U, \bar{S}$.

Output: $S_{U \cup \Delta U}$ on $U \cup \Delta U$.

Step 1: Compute ΔI_S .

Step 2: If $|\Delta I_S| = 0$, go to Step 6, else go to Step 4.

Step 3: Compute $I(Y; S), H(Y)$ on ΔI_S .

If $I(Y; S) = H(Y)$, go to Step 6;

Else go to Step 4.

Step 4: While $I(Y; S) \neq H(Y)$ do.

{
 For every $k \in \bar{S}$, compute $Sig(x_k, S, Y)$ on ΔI_S ;
 Select $k^* = \arg_k \max \{Sig(x_k, S, Y)\}$;
 $S \leftarrow S \cup \{k^*\}$, $\bar{S} \leftarrow \bar{S} - \{k^*\}$;
 Update $I(Y; S)$ on ΔI_S .
 }

Step 5: $S_{U \cup \Delta U} \leftarrow S$.

Step 6: Return $S_{U \cup \Delta U}$.

Table 1: Time complexity of the algorithm

Algorithm	Time Complexity
mRMR	$O(U + \Delta U \log U + \Delta U)$
I-mRMR	$\max\{O(S U + \Delta U), O(\Delta I_S (\log \Delta I_S) F - S S)\}$

4 NUMERICAL EXPERIMENTS

In this section, we conduct some numerical experiments to evaluate the proposed algorithm, I-mRMR, on ten datasets from UCI and two extremely high dimensional datasets from KDD cup, seen in Table 2&8. The Max-Relevance and Min-Redundancy feature selection based on normalized mutual information, denoted by mRMR [9], as the classical non-incremental feature selection algorithm, is compared with I-mRMR. Also, an entropy-based incremental feature selection algorithm GIARC, as an incremental feature selection algorithm, is compared with our proposed algorithm [21]. Because GIARC and I-mRMR are comparable as both of their measures of feature relevance are generalized by information entropy.

4.1 Experimental Setup

All the experiments have been conducted on computer with CentOS release 6.5(Final), Westmere E56xx/L56xx/X56xx(Nehalem-C) and 8GB memory. The programming language is Python. The detail experimental setting are presented as follows.

(1) Since our algorithm is only valid for discrete data, fuzzy-c-means is used to discretize those continuous data sets.


(2) Every dataset is divided into six parts equally, the first part is used as the original data set U , and remaining parts as the newly added dataset ΔU , are added one by one.

(3) All the experimental comparison is demonstrated from four indices: running time, global speedup ratio, local speedup ratio and classification accuracy.

Global speedup ratio: $\frac{\sum_{streaming\ instances} RT_{mRMR}}{\sum_{streaming\ instances} RT_{I-mRMR}}$

Where RT_{mRMR} denotes the running time of mRMR on the seen instances so far, RT_{I-mRMR} denotes the running time of I-mRMR on the seen instances so far. When the dataset is divided into six

Table 2: The description of the selected datasets from UCI

Dataset	Instances	Features	Data Type	Classes
madelon	2000	500	continuous	2
colon	62	2000	continuous	2
breast	84	9216	continuous	5
arcene	100	10000	continuous	2
Gene9	203	12600	continuous	5
TCGA_PANCAN	801	6383	continuous	5
Ad	2359	1558	continuous	2
FPS5	3600	3208	continuous	6
 Gisette	3600	4812	continuous	6
Gisette	7000	5000	continuous	2

parts, $\sum_{streaming\ instances} RT_{mRMR}$ represents the sum of six times running time of mRMR, where each time the dataset is updated when some new instances arriving.

Local speedup ratio: $\frac{RT_{mRMR}}{RT_{I-mRMR}}$

When the dataset is divided into six parts, the local speedup ratio is the ratio of the running time of mRMR on the whole dataset to the running time of I-mRMR when the last part arriving.

(4) To show the effectiveness of I-mRMR, SVM and KNN are used to evaluate the classification performance. And 5-fold cross validation is used in classification evaluation.

4.2 Experimental I: Evaluation on UCI

To test the performance of I-mRMR, some experimental comparison and analyses are conducted on ten UCI datasets, seen in Table 2. Among the datasets in Table 2, 'madelon', 'arcene' are used in the NIPS 2003 Feature selection challenge. The dataset 'colon' is frequently studied as public microarray datasets [31].

4.2.1 Compared with mRMR. In this part, I-mRMR and mRMR are compared. Both of them are feature selection methods based on normalized mutual information of "Max-Relevance and Min-Redundancy" criterion. One main difference between them is that I-mRMR is an incremental feature selection algorithm, whereas mRMR is a non-incremental feature selection algorithm. Another obvious difference of them lies in the different stop criteria. The stopping criterion of I-mRMR is that the mutual information of the selected feature subset is equal to the information entropy of the label feature Y , whereas the stop criterion of mRMR is to select top k features.

A. The Comparison of the running time

In this part, we demonstrate the running time of I-mRMR and mRMR when instances successively arriving and then graph them in Figure 1.

Figure 1 clearly demonstrates that the running time of I-mRMR changes slightly, whereas the running time of mRMR increases significantly with the instances successively arriving. This shows that I-mRMR works efficiently on streaming instances, whereas mRMR works more and more less-efficiently.

To further illustrate the time superiority of I-mRMR, the global speedup ratio is then presented, seen in Tables 3&4.

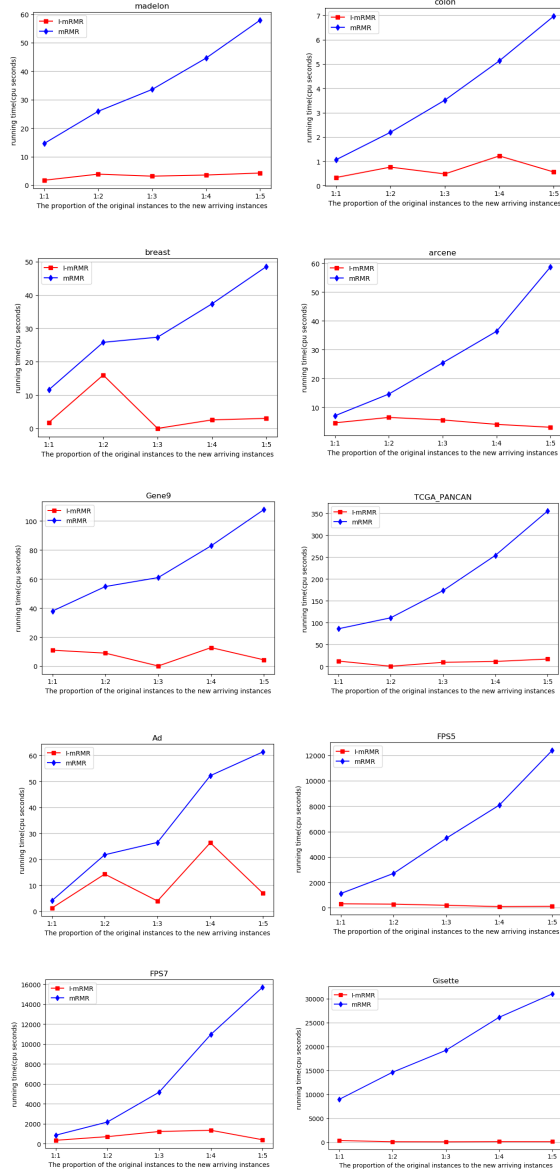


Figure 1: The running time of I-mRMR and mRMR with instances successively arriving

Table 3 shows that the total time of mRMR is obviously or even significantly higher than that of I-mRMR, especially on the datasets with high number of instances. This is because when some new instances arriving mRMR has to be recomputed on the whole seen instances so far, which is really time consuming. Take 'Gisette' as an example, we present the detail running time of I-mRMR and mRMR in Table 4. Table 4 clearly demonstrates that mRMR is really time consuming and conducts much redundant computation. Furthermore, Table 5 demonstrates the time superiority of I-mRMR from the aspect of local speedup ratio. From Table 5 we observe that I-mRMR is significantly or even dramatically faster than mRMR.

Table 3: The global speedup ratio of mRMR and I-mRMR

Dataset	I-mRMR	mRMR	Global speedup ratio
madelon	184.32s	32.49s	5.67
colon	19.08s	3.79s	5.03
breast	156.96s	29.52s	5.31
arcene	144.37s	29.73s	4.85
Gene9	377.17s	69.47s	5.43
TCGA_PANCAN	1026.37s	121.77s	8.42
Ad	168.14s	58.17s	2.89
FPS5	30092s (8hr21m32s)	1654s (27m34s)	18.19
FPS7	35092s (9hr44m52s)	4501s (1hr25m1s)	7.79
Gisette	103161s (28hr39m21s)	10801s (3hr1s)	9.55
Average	17420s (4hr50m6s)	1730s(28m50s)	7.31

This is because I-mRMR only consider part of instances which are not distinguished by the previous selected features, whereas mRMR computes on the whole seen instances so far. Take three datasets with high number of instances 'FPS5', 'FPS7' and 'Gisette' as examples. We find that the running time of I-mRMR when the 5th part of new instances arriving is far less than that of mRMR. This is because I-mRMR find that the previously selected features are enough to distinguish the new arriving instances and no additive feature need to be added, which dramatically saves the running time and avoids the redundant computation.

B. The Comparison of the classification accuracy

In this part, to illustrate the effectiveness of I-mRMR, we compare the classification performance of I-mRMR and mRMR. For fairness, the number of the features selected by mRMR is set as same as the number of selected features by I-mRMR. The Classification accuracies of mRMR and I-mRMR are summarized in Table 6.

Several observations can be drawn from the results in Table 6. First, I-mRMR and mRMR have the same number of selected features. This ensures the comparability of these two algorithms. Second, we find that I-mRMR has the similar or sometime slightly lower classification accuracy than mRMR. Whereas, I-mRMR has a better classification performance with dramatical dimension reduction than the original datasets. All these indicate that I-mRMR has an acceptable and effective classification performance.

4.2.2 Compared with GIARC. In this part, we compare I-mRMR with GIARC. One of the main similarities is that both of them utilize the measure generalized by information entropy to select the representative feature, another similarity is that both of them are incremental feature selection methods. As a result, it is necessary to compare their performance.

For fairness, the cost of de-redundancy of GIARC is not counted. That is to say, the real running time of GIARC is higher than presented in this part.

Table 4: the detail running time of Gisette when new instances arriving successively

Algorithm		1st new instances arriving	2st new instances arriving	3st new instances arriving	4st new instances arriving	5st new instances arriving
I-mRMR	Selected Feature No.	69	73	74	76	76
	running time	337s	100.85s	74.6s	119s	105.6s
mRMR	Selected Feature No.	69	73	74	76	76
	running time	8954s(2hr29m14s)	14617s(4hr3m23s)	19188s(5hr19m48s)	26088s(7hr14m48s)	30991s(8hr36m31s)

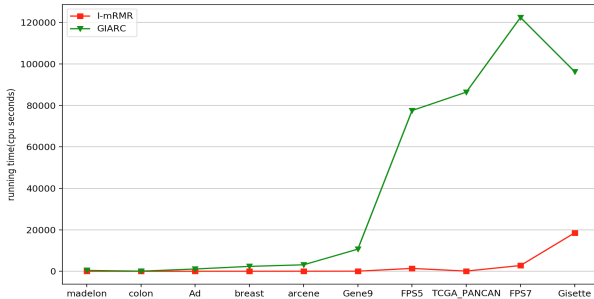
Table 5: The local speedup ratio of mRMR and I-mRMR

Dataset	mRMR	I-mRMR	Local speedup ratio
madelon	57.92s	4.23s	13.69
colon	19.08s	0.56	34.07
breast	48.53s	3.02s	16.06
arcene	58.7s	3.05s	19.24
Gene9	107.9s	4.25s	25.38
TCGA_PANCAN	355.8s	16.74s	21.25
Ad	61.36s	6.88s	8.92
FPS5	12392s (3hr26m32s)	119s	104
FPS7	15690s (4hr21m30s)	398s	39.4
Gisettee	30991s (8hr36m31s)	105.6s	293.5
Average	5978s (1hr39m38s)	66s	57.5

In the following, the total running time of these two algorithms are compared and graphed in Figure 2. Here, the total running time means the sum of time running mRMR/I-mRMR six times, where each time the dataset is updated when some new instances arriving.

A. The comparison of total running time

The total running time of I-mRMR compared with GIARC and graphed in Figure 2.

**Figure 2: The total running time of I-mRMR and GIARC**

The classification accuracy of features selected by I-mRMR and GIARC is shown in Table 7.

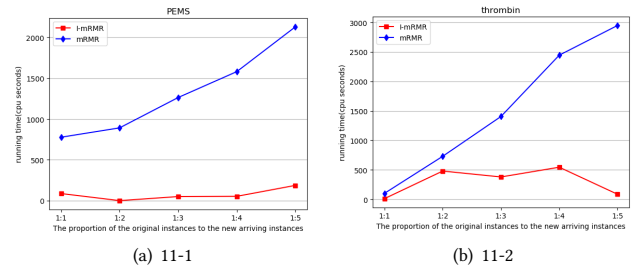
According to the experimental results of this part, we can see from Figure 2 that GIARC consumes more time than I-mRMR, especially on the datasets with both high number of instances and high dimensionality, e.g. FPS5, FPS7, TCGA_PAVCAN and Gisette. In addition, it is easy from Table 7 to see that the classification accuracy of I-mRMR on all the datasets is obviously higher than that of GIARC. Although GIARC could select some more features to increase its classification accuracy by adjusting its threshold, it makes GIARC cost some more time and is far time consuming than I-mRMR. It is easy to get that I-mRMR is more feasible and effective.

4.3 Experimental II: Evaluation on High dimensional Data Set

In this section, we conduct numerical experiments on two extremely high dimensional datasets, seen in Table 8, Here 'PEMS' comes from the California Department of Transportation PEMS website¹, and 'thrombin' comes from KDD Cup 2001 competition.

A. The comparison of the running time

First, we compare the running time of mRMR and I-mRMR on the two extremely high dimension datasets. Because GIARC spends too much time on these datasets, we have to terminate them when they run more than one week. The trends of the running time of I-mRMR and mRMR with the instances successively arriving are graphed in Figure 3 and the global/local speedup ratio of them are presented in Table 9&10.

**Figure 3: The trends of the running time of I-mRMR and mRMR with the instances successively arriving**

¹www.pems.dot.gov

Table 6: The classification accuracy of mRMR and I-mRMR

Dataset	Original			mRMR			I-mRMR		
	Feature No.	Accuracy		Feature No.	Accuracy		Feature No.	Accuracy	
		SVM	KNN		SVM	KNN		SVM	KNN
madelon	500	0.573	0.585	15	0.614	0.577	15	0.581	0.567
colon	2000	0.803	0.791	6	0.835	0.834	6	0.831	0.832
breast	9216	0.829	0.746	7	0.865	0.893	7	0.777	0.754
arcene	10000	0.679	0.649	7	0.699	0.661	7	0.699	0.667
Gene9	12600	0.800	0.803	8	0.889	0.876	8	0.906	0.895
TCGA-PANCAN	16383	0.996	0.998	8	0.976	0.983	8	0.985	0.976
Ad	1558	0.839	0.947	7	0.868	0.924	7	0.943	0.947
FPS5	3208	0.748	0.535	82	0.797	0.721	82	0.793	0.744
FPS7	4812	0.816	0.596	79	0.777	0.678	79	0.784	0.726
Gisette	5000	0.896	0.963	76	0.943	0.932	76	0.946	0.951
Average	645.9	0.798	0.761	29.5	0.826	0.808	29.5	0.825	0.806

Table 7: Classification accuracy of I-mRMR and GIARC

Dataset	I-mRMR			GIARC		
	Feature No.	Accuracy		Feature No.	Accuracy	
		SVM	KNN		SVM	KNN
madelon	15	0.581	0.567	9	0.507	0.568
colon	6	0.831	0.832	4	0.823	0.76
breast	7	0.777	0.754	5	0.5	0.488
arcene	7	0.699	0.667	6	0.698	0.662
Gene9	8	0.906	0.895	6	0.716	0.814
TCGA-PANCAN	8	0.985	0.976	8	0.918	0.899
Ad	7	0.943	0.947	5	0.914	0.914
FPS5	82	0.793	0.744	42	0.603	0.575
FPS7	79	0.784	0.726	32	0.502	0.491
Gisette	76	0.946	0.951	10	0.884	0.865
Average	29.5	0.825	0.806	12.7	0.707	0.704

Table 8: Summary of the two high dimensional datasets.

Dataset	Instances	Features	Data Type	Classes
PEMS	267	138672	Continuous	7
thrombin	1909	139351	Discrete	2

Table 9: The Global speedup ratio of I-mRMR and mRMR

Dataset	mRMR	I-mRMR	Global speedup Ratio
PEMS	7172s (1hr59m28s)	1287s(21m27s)	5.57
thrombin	7680s (2hr40m)	1872s(31m12s)	4.1
Average	7426s (2hr3m46s)	1579.5s (26m19.5s)	4.48

Table 10: Local Speedup Ratio of mRMR and I-mRMR

Dataset	mRMR	I-mRMR	Local speedup Ratio
PEMS	2130s (35m30s)	187s	6.37
thrombin	2949s (49m9s)	85s	34.7
Average	2539.5s (42m19.5s)	136s	23.02

Several observations could be obtained from the above figures and tables. First, from Figure 3 we find that mRMR is more time consuming than I-mRMR with the instances successively arriving. Second, Table 9&10 globally and locally demonstrate that I-mRMR works significantly more efficient than mRMR on the datasets with extremely high dimensionality.

B. The comparison of classification accuracy

The classification accuracy of I-mRMR and mRMR on extremely high-dimensional datasets is shown in Table 11.

Table 11 shows the classification accuracy of mRMR and I-mRMR on the two high-dimensional data sets. The average classification accuracy of I-mRMR is similar or sometimes even higher than the average classification accuracy of mRMR. Therefore, these results verify that the proposed incremental feature selection algorithm is feasible and efficient on the extremely high-dimensional datasets with an acceptable classification accuracy.

In summary, this numerical experimental part demonstrates that I-mRMR is more efficient than mRMR with the acceptable classification performance.

5 CONCLUSIONS

In this paper, we propose an incremental feature selection algorithm I-mRMR based on max-relevance and min-redundancy criterion. When a new set of instances is arriving, not all seen instances so far are necessary to update the feature selection results. Actually, just an Incremental Key Instance Set, which is composed of

Table 11: The classification accuracy of I-mRMR and mRMR

Dataset	Original			mRMR			I-mRMR		
	Feature No.	Accuracy		Feature No.	Accuracy		Feature No.	Accuracy	
		SVM	KNN		SVM	KNN		SVM	KNN
PEMS	138672	0.869	0.787	10	0.829	0.865	10	0.855	0.904
thrombin	139351	0.978	0.982	5	0.978	0.973	5	0.983	0.985
Average	139011	0.924	0.885	7.5	0.904	0.919	7.5	0.919	0.945

the instances undistinguished by historical selected features, is key to update the feature subset. As a result, I-mRMR is designed by using Incremental Key Instance Set, which dramatically improve the efficiency of feature selection on streaming instances. By numerical experiments, we demonstrate that the proposed incremental algorithm is significantly faster than the classical algorithm mRMR not only in the global speedup ratio but also in the local speedup ratio. Furthermore, on the extremely high-dimensional dataset, we experimentally demonstrate that our proposed feature selection algorithm I-mRMR is obviously more efficient than mRMR with an acceptable classification accuracy.

ACKNOWLEDGMENTS

REFERENCES

- [1] Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *Neural Networks IEEE Transactions on* 5, 4 (1994), 537–550.
- [2] David A. Bell and Hui Wang. 2000. A Formalism for Relevance and Its Application in Feature Subset Selection. *Machine Learning* 41, 2 (01 Nov 2000), 175–195. <https://doi.org/10.1023/A:1007612503587>
- [3] L. Bruzzone and D. Fernández Prieto. 1999. An incremental-learning neural network for the classification of remote-sensing images. *Pattern Recognition Letters* 20, 11–13 (1999), 1241–1248.
- [4] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen. 1992. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks* 3, 5 (1992), 698–713.
- [5] Girish Chandrashekar and Ferat Sahin. 2014. *A survey on feature selection methods*. Pergamon Press, Inc. 16–28 pages.
- [6] T. W. S Chow and D Huang. 2005. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural Networks* 16, 1 (2005), 213–224.
- [7] C. Ding and H. Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics & Computational Biology* 3, 02 (2005), 185–205.
- [8] Gauthier Doquire and Michel Verleysen. 2011. *Feature Selection with Mutual Information for Uncertain Data*. Springer Berlin Heidelberg. 330–341 pages.
- [9] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada. 2009. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks* 20, 2 (2009), 189–201.
- [10] Wenfei Fan, Jianzhong Li, Nan Tang, and Wenyuan Yu. 2012. Incremental Detection of Inconsistencies in Distributed Data. *IEEE Transactions on Knowledge & Data Engineering* 26, 6 (2012), 1–1.
- [11] D. François, F. Rossi, V. Wertz, and M. Verleysen. 2007. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing* 70, 7–9 (2007), 1276–1288.
- [12] João Gama, André Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A Survey on Concept Drift Adaptation. *ACM Comput. Surv.* 46, 4, Article 44 (March 2014), 37 pages. <https://doi.org/10.1145/2523813>
- [13] Y. F. Gao, B. Q. Li, Y. D. Cai, K. Y. Feng, Z. D. Li, and Y. Jiang. 2013. Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. *Molecular Biosystems* 9, 1 (2013), 61–69.
- [14] Isabelle Guyon, André Elisseeff, and Leslie Pack Kaelbling. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 6 (2003), 1157–1182.
- [15] Feng Hu, Guoyin Wang, Hai Huang, and Yu Wu. 2005. Incremental Attribute Reduction Based on Elementary Sets. 36, 41 (2005), 185–193.
- [16] Xuegang Hu, Peng Zhou, Peipei Li, Jing Wang, and Xindong Wu. 2017. A survey on online feature selection with streaming features. *Frontiers of Computer Science* 1 (2017), 1–15.
- [17] Jinjie Huang, Yunze Cai, and Xiaoming Xu. 2006. A Wrapper for Feature Selection Based on Mutual Information. In *International Conference on Pattern Recognition*. 618–621.
- [18] Dimitrios Kales and Tim Morris. 1996. Efficient incremental induction of decision trees. *Machine Learning* 24, 3 (01 Sep 1996), 231–242. <https://doi.org/10.1007/BF00058613>
- [19] Ivan Kojadinovi. 2009. Relevance measures for subset variable selection in regression problems based on k-additive mutual information. *Computational Statistics & Data Analysis* 49, 4 (2009), 1205–1227.
- [20] N Kwak and C. H. Choi. 2002. Input feature selection for classification problems. *IEEE Transactions on Neural Networks* 13, 1 (2002), 143.
- [21] Jiye Liang, Feng Wang, Chuangyin Dang, and Yuhua Qian. 2013. A Group Incremental Approach to Feature Selection Applying Rough Set Technique. *IEEE Transactions on Knowledge & Data Engineering* 26, 2 (2013), 294–308.
- [22] H. Peng, F. Long, and C. Ding. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 27, 8 (2005), 1226–1238.
- [23] R Polikar, J Byorick, S Krause, and A Marino. 2002. Learn++: a classifier independent incremental learning algorithm for supervised neural networks. In *International Joint Conference on Neural Networks*. 1742–1747.
- [24] F. Rossi, A. Lendasse, D. François, V. Wertz, and M. Verleysen. 2006. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics & Intelligent Laboratory Systems* 80, 2 (2006), 215–226.
- [25] Donald L. Schilling. 2003. *Elements of Information Theory*. Wiley. 155–183 pages.
- [26] Jeffrey C. Schlimmer and Douglas H. Fisher. 1986. A Case Study of Incremental Concept Induction. In *National Conference on Artificial Intelligence*. Philadelphia, Pa, August 11–15, 1986. Volume 1: Science. 496–501.
- [27] Liu Zong Tian. 1999. An Incremental Arithmetic for the Smallest Reduction of Attributes. *Acta Electronica Sinica* (1999).
- [28] Jorge R. Vergara and Pablo A. Estévez. 2014. A review of feature selection methods based on mutual information. *Neural Computing & Applications* 24, 1 (2014), 175–186.
- [29] Jie Xu, Chen Xu, Bin Zou, Yuan Yan Tang, Jiangtao Peng, and Xinge You. 2018. New Incremental Learning Algorithm With Support Vector Machines. *IEEE Transactions on Systems Man & Cybernetics Systems* PP, 99 (2018), 1–12.
- [30] Jieping Ye, Qi Li, Hui Xiong, Haesun Park, Ravi Janardan, and Vipin Kumar. 2005. IDR/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition. *IEEE Transactions on Knowledge & Data Engineering* 17, 9 (2005), 1208–1222.
- [31] Lei Yu, Chris Ding, and Steven Loscalzo. 2008. Stable feature selection via dense feature groups. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, Usa, August*. 803–811.
- [32] Lei Yu and Huan Liu. 2004. Efficient Feature Selection Via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research* 5, 12 (2004), 1205–1224.
- [33] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu. 2015. i^2 MapReduce: Incremental MapReduce for Mining Evolving Big Data. *IEEE Transactions on Knowledge & Data Engineering* 27, 7 (2015), 1906–1919.
- [34] Lei Zhu, Shaoning Pang, Abdolhossein Sarrafzadeh, Tao Ban, and Daisuke Inoue. 2016. Incremental and Decremental Max-Flow for Online Semi-Supervised Learning. *IEEE Transactions on Knowledge & Data Engineering* 28, 8 (2016), 2115–2127.