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

Yeliang Xiu[1], Suyun Zhao[1,2(✉)], Hong Chen[1,2], and Cuiping Li[1,2]

[1] Information of School, Renmin University of China,
Beijing, People's Republic of China
zhao.suyun@yahoo.com, 2017104068@ruc.edu.cn
[2] Key Laboratory of Data Engineering and Knowledge Engineering,
Ministry of Education, Beijing, China
http://deke.ruc.edu.cn

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 Incremental Key Instance Set is far less than the whole instances, the incremental feature selection by using this key set avoids 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.

Keywords: Feature selection · Incremental algorithm · Normalized mutual information · Min-Redundancy · Max-Relevance

1 Introduction

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 [4,13]. The main idea of incremental feature selection is that only part of the data are 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 [6,12].

Feature selection based on mutual information has been deeply studied [2,10,11,14], because mutual information (MI) [8] is a good tool to measure the correlation and redundancy among features. As a pioneer, Battiti [1] proposed

J. Shao et al. (Eds.): APWeb-WAIM 2019, LNCS 11642, pp. 103–110, 2019.
https://doi.org/10.1007/978-3-030-26075-0_8

a greedy selection method called MIFS based on mutual information between inputs and outputs. Considering MIFS does not work well on nonlinear problems, Kwak and Choi [5] 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. [7] then proposed a heuristic "Max-Relevance and Min-Redundancy" framework for feature selection. In [7] it is pointed that mRMR criterion is equal to max-dependency. Furthermore, Estévez and Tesmer [3] proposed an updated feature selection method, called normalized mutual information features selection. However, 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.

In this paper, we propose an incremental feature selection algorithm, called I-mRMR. First, Incremental Key Instance Set is proposed which is composed of the instances not distinguished by historical selected features. An incremental algorithm is then proposed based on this Incremental Key Instance Set. Finally, the numerical experiments of I-mRMR shows that I-mRMR 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 mRMR based on the normalized mutual information. Section 3 introduces the concept of Incremental Key Instance Set and presents the incremental feature selection algorithm, I-mRMR. In Sect. 4, ten UCI datasets are employed to illustrate the effectiveness and efficiency of I-mRMR. Section 5 concludes this paper.

2 Preliminaries

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

2.1 Notation Description

Given a set of original instances $U = [x^{(1)}, x^{(2)}, \cdots, x^{(n)}]^T$. 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. $\overline{S}$ 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 $S(i = 1, \cdots, n)$, $Y = [y^{(1)}, \cdots .y^{(n)}]^T$ is a column vector representing the label feature in U. Here $y^{(i)}$ is the label for the i-th instance in $U(i = 1, \cdots, n)$.

2.2 Max-Relevance and Min-Redundancy

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

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

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.

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

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

$$max\ \Phi(D, R), \Phi = D - R \tag{3}$$

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:

$$max_{k \in F - S_{m-1}} [I(Y, x_k) - \frac{1}{|S_{m-1}|} \sum_{t \in S_{m-1}} I(x_k, x_t)] \tag{4}$$

2.3 Normalized Mutual Information Feature Selection

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

$$NI(x_k, x_t) = \frac{I(x_k, x_t)}{min\{H(x_k), H(x_t)\}} \tag{5}$$

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

$$max_k [I(Y, x_k) - \frac{1}{|S_{m-1}|} \sum_{t \in S_{m-1}} NI(x_k, x_t)] \tag{6}$$

3 The Proposed Incremental Algorithm

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.

3.1 Problem Definition

When some new instances, represented by $\triangle U \in R^{m \times p}$(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 $\triangle U$, $j = 1, \cdots, m$. The selected feature subset S has to be updated from U to $U \cup \triangle U$. The traditional method is directly to recompute the feature selection method on all seen instances $U \cup \triangle U$ to obtain the updated feature subset $S_{U \cup \triangle U}$. 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.

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.

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 $\triangle U$, then Incremental Key Instance Set of S, denoted by $\triangle I_S$, is defined as follows.*

$$\begin{aligned} \triangle I_S = \{x^{(i)} \in U | \exists x^{(n+j)} \ s.t. \ x_S^{(i)} = x_S^{(n+j)}, y^{(i)} \neq y^{(n+j)}\} \cup \\ \{x^{(n+j)} \in \triangle U | \exists x^{(i)} \in U \ s.t. \ x_S^{(i)} = x_S^{(n+j)}, y^{(i)} \neq y^{(n+j)}\} \end{aligned} \tag{7}$$

Incremental Key Instance Set $\triangle I_S$ 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. $\triangle I_S$ 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.

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

$$Sig(x_k, S, Y) = I(Y, x_k) - \frac{1}{|S|} \sum_{t \in S} NI(x_k, x_t) \tag{8}$$

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

3.3 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 1.

4 Numerical Experiments

In this section, we conduct some numerical experiments to evaluate the proposed algorithm, I-mRMR, on ten datasets from UCI. The Max-Relevance and Min-Redundancy feature selection based on normalized mutual information, denoted by mRMR [3], as the classical non-incremental feature selection algorithm, is compared with I-mRMR.

Algorithm 1: An incremental algorithm for feature selection based on Max-Relevance, and Min-Redundancy (I-mRMR)

```
Input: $U$, $F$, $Y$, $S$, $\triangle U$, $\overline{S}$.
Output: $S_{U \cup \triangle U}$ on $U \cup \triangle U$.
Step 1: Compute $\triangle I_S$.
Step 2: If $| \triangle I_S|$=0, go to Step 6, else go to Step 4.
Step 3: Compute $I(Y;S)$,$H(Y)$ on $\triangle I_S$.
        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 $k \in \overline{S}$, compute $Sig(x_k, S, Y)$ on $\triangle I_S$;
        Select $k^* = arg_k max\{Sig(x_k, S, Y)\}$;
        $S \leftarrow S \cup \{k^*\}$, $\overline{S} \leftarrow \overline{S} - \{k^*\}$;
        Update $I(Y;S)$ on $\triangle I_S$.
    }
Step 5: $S_{U \cup \triangle U} \leftarrow S$.
Step 6: Return $S_{U \cup \triangle U}$.
```

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 8 GB 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 $\triangle U$, are added one by one.

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

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 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: Evaluation on UCI

To test the performance of I-mRMR, some experimental comparison and analyses are conducted on ten UCI datasets.

Compared with mRMR. In this part, I-mRMR and mRMR are compared. Both of them are feature selection methods based on the 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.

we demonstrate the running time of I-mRMR and mRMR when instances successively arriving and then graph them in Fig. 1.

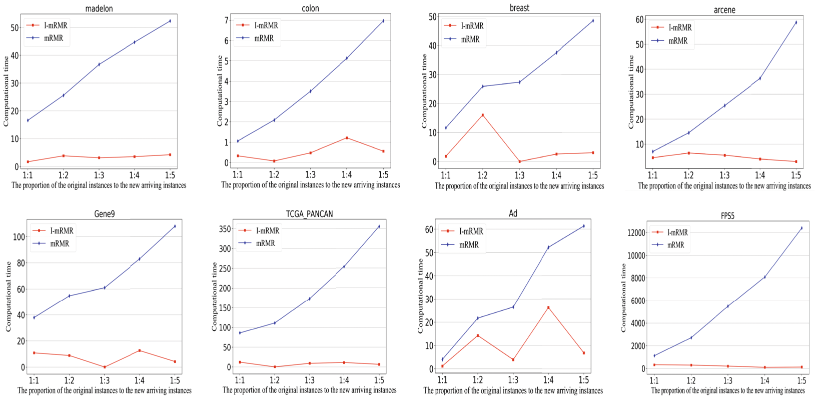

Fig. 1. The running time of I-mRMR and mRMR with instances successively arriving

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 Table 1.

Table 1 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. Furthermore, Table 2 demonstrates the time superiority of I-mRMR from the aspect of local speedup ratio. From Table 2 we observe that I-mRMR is significantly or even dramatically faster than mRMR. 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.

Table 1. The global speedup ratio of mRMR and I-mRMR

Dataset	mRMR	I-mRMR	Global speedup ratio
madelon	184.32 s	32.49 s	5.67
colon	19.08 s	3.79 s	5.03
breast	156.96 s	29.52 s	5.31
arcene	144.37 s	29.73 s	4.85
Gene9	377.17 s	69.47 s	5.43
TCGA_PANCAN	1026.37 s	121.77 s	8.42
Ad	168.14 s	58.17 s	2.89
FPS5	30092 s (8 h 21 m 32 s)	1654 s (27 m 34 s)	18.19
FPS7	35092 s (9 h 44 m 52 s)	4501 s (1 h 25 m 1 s)	7.79
Gisette	103161 s (28 h 39 m 21 s)	10801 s (3 h 1 s)	9.55
Average	17420 s (4 h 50 m 6 s)	1730 s (28 m 50 s)	7.31

Table 2. The local speedup ratio of mRMR and I-mRMR

Dataset	mRMR	I-mRMR	Local speedup ratio
madelon	57.92 s	4.23 s	13.69
colon	19.08 s	0.56	34.07
breast	48.53 s	3.02s	16.06
arcene	58.7 s	3.05 s	19.24
Gene9	107.9 s	4.25 s	25.38
TCGA_PANCAN	355.8 s	16.74 s	21.25
Ad	61.36 s	6.88 s	8.92
FPS5	12392 s (3 h 26 m 32 s)	119 s	104
FPS7	15690 s (4 h 21 m 30 s)	398 s	39.4
Gisettee	30991 s (8 h 36 m 31 s)	105.6 s	293.5
Average	5978 s (1 h 39 m 38 s)	66 s	57.5

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 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.

Acknowledgements. This work is supported by National Key Research & Develop Plan (No. 2016YFB1000702, 2018YFB1004401), National Key R&D Program of China(2017YFB1400700), NSFC under the grant No. 61732006, 61532021, 61772536, 61772537, 61702522 and NSSFC (No. 12&ZD220), and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (15XNLQ06). It was partially done when the authors worked in SA Center for Big Data Research in RUC. This Center is funded by a Chinese National 111 Project Attracting. This work is also supported by the Macao Science and Technology Development Fund (081/2015/A3).

References

1. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**(4), 537–550 (1994)
2. Chandrashekar, G., Sahin, F.: A Survey on Feature Selection Methods. Pergamon Press, Inc., Oxford (2014)
3. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. IEEE Trans. Neural Netw. **20**(2), 189–201 (2009)
4. Guyon, I., Elisseeff, A., Kaelbling, L.P.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(6), 1157–1182 (2003)
5. Kwak, N., Choi, C.H.: Input feature selection for classification problems. IEEE Trans. Neural Netw. **13**(1), 143 (2002)
6. Liu, H., Setiono, R.: Incremental feature selection. Appl. Intell. **9**(3), 217–230 (1998). https://doi.org/10.1023/A:1008363719778
7. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
8. Rossi, F., Lendasse, A., François, D., Wertz, V., Verleysen, M.: Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. Chemom. Intell. Lab. Syst. **80**(2), 215–226 (2006)
9. Schilling, D.L.: Elements of Information Theory. Wiley, Hoboken (2003)
10. Sluga, D., Lotrič, U.: Quadratic mutual information feature selection. Entropy **19**(4), 157 (2017)
11. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**(1), 175–186 (2014)
12. Xu, J., Xu, C., Zou, B., Tang, Y.Y., Peng, J., You, X.: New incremental learning algorithm with support vector machines. IEEE Trans. Syst. Man Cybern. Syst. **PP**(99), 1–12 (2018)
13. Ye, J., Li, Q., Xiong, H., Park, H., Janardan, R., Kumar, V.: IDR/QR: an incremental dimension reduction algorithm via QR decomposition. IEEE Trans. Knowl. Data Eng. **17**(9), 1208–1222 (2005)
14. Zhang, Z., Hancock, E.R.: Mutual information criteria for feature selection. In: Pelillo, M., Hancock, E.R. (eds.) SIMBAD 2011. LNCS, vol. 7005, pp. 235–249. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24471-1_17