Forward Feature Selection Based on Approximate Markov Blanket

Min Han and Xiaoxin Liu

Faculty of Electronic Information and Electrical Engineering,
Dalian University of Technology, Dalian, China
minhan@dlut.edu.cn, xiaoxinliu@mail.dlut.edu.cn

Abstract. Feature selection has many applications in solving the problems of multivariate time series . A novel forward feature selection method is proposed based on approximate Markov blanket. The relevant features are selected according to the mutual information between the features and the output. To identify the redundant features, a heuristic method is proposed to approximate Markov blanket. A redundant feature is identified according to whether there is a Markov blanket for it in the selected feature subset or not.The simulations based on the Friedman data, the Lorenz time series and the Gas Furnace time series show the validity of our proposed feature selection method.

Keywords: Feature Selection, Redundancy Analysis, Markov Blanket, Mutual Information.

1 Introduction

Feature selection is very important in the modeling of multivariate time series. There are three advantages of feature selection [1]. Firstly, with feature selection the forecasting or classification accuracy can be improved. Secondly, the time and storage cost can be reduced. Thirdly, a better understanding of the underlying process that generated the data can be obtained.

Feature selection refers to methods that select the best subset of the original feature set. The best subset is supposed to contain all the relevant features and get rid of all the irrelevant and redundant features. Mutual information (MI) is a commonly used criterion for the correlation of features. The MI measures the amount of information contained in a feature or a group of features, in order to predict the dependent one. It can not only capture the linear correlation between features, but also capture the nonlinear correlation. Additionally, it has no assumption on the distribution of the data. So we apply MI as the criterion of feature selection in this paper.

Battiti proposed a mutual information feature selector (MIFS) which selects the feature that maximizes MI between feature and the output, corrected by subtracting a quantity proportional to the sum of MI with the previously selected features [2]. But it depends on the proportion parameter to determine whether the feature is redundant or not. A variant of MIFS which can overcome this disadvantage is min-redundancy max-relevance (mRMR) criterion [3]. It maximizes MI between feature and the

J. Wang, G.G. Yen, and M.M. Polycarpou (Eds.): ISNN 2012, Part II, LNCS 7368, pp. 64–72, 2012.

output, and minimizes the average MI between feature with the selected ones. It has a better performance than MIFS, but it tends to select features which have more values. To solve this problem, Estévez proposed a method named normalized mutual information feature selection (NMIFS) which normalizes the MI by the minimum entropy of both features [4].

Because the computation cost of high-dimensional MI estimation is usually very high. The above methods are all incremental search schemes that select one feature at a time. At each iteration, a certain criterion is maximized with respect to a single feature, not taking into account the interaction between groups of features. This may cause the selection of redundant features. Besides, all the methods will not stop until the desired number of features are selected. Yu and Liu proposed a fast correlation-based filter (FCBF) based on Markov blanket [5]. The method not only focuses on finding relevant features, but also performs explicit redundancy analysis.

Markov blanket was proposed by Koller and Sahami [6]. It is pointed out that an optimal subset can be obtained by a backward elimination procedure, known as Markov blanket filtering. In this paper, we propose a forward feature selection method based on approximate Markov blanket. The MI is estimated by the method of *k* nearest neighbors (*k*-NN) [7] which is easier and faster than the kernel methods, and can estimate the high dimensional MI. The MI between input and output is used as relevant criterion and the Markov blanket is used as redundant criterion. While in [8] only the relevancy is measured and the Markov blanket is used as the stopping criterion. Simulation results substantiate the proposed method on both artificial and benchmark datasets.

2 Markov Blanket

If X and Y is continuous random features with probability density function $p(x)$ and $p(y)$, and the joint probability density function between them is $p(x,y)$, then the MI between X and Y is

$$I(X;Y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad . \tag{1}$$

Let F be the set of all features, X_i is one of it and Y is the output. Let $Z=\{F,Y\}$ and $I(X,Y)$ represents the MI between X and Y. The Markov blanket of X_i can be defined as follows.

Definition 1 (Markov Blanket). Given a feature X_i, let $M_i \subset F(X_i \notin M_i)$, M_i is said to be a Markov blanket for X_i if

$$I(\{M_i \cup X_i\},Z-\{M_i \cup X_i\}) \approx I(M_i,Z-\{M_i \cup X_i\}) \quad . \tag{2}$$

According to this definition, if M_i is the Markov blanket for X_i, it subsumes all the information that X_i has about *Z*. That is to say, as for *Z*, X_i is redundant given the subset M_i. And it leads to the following corollary [8].

Corollary 1. Let $S \subset F(X_i \notin S)$, if in $Z = \{F, Y\}$, $M_i \subset S$ and it is the Markov blanket for X_i, then $I(S,Y) \approx I(S \cup \{X_i\}, Y)$.

Therefore, if we can find a Markov blanket for X_i in the selected feature subset *S* during the forward selection, the correlation between X_i and *Y* can be totally replaced by *S*, which means that X_i is redundant for *S* and it should be removed.

3 Feature Selection Based on Markov Blanket

Although Markov blanket can measure the redundancy between features, the process of searching for a Markov blanket is an exhausting process which is somewhat like feature selection. With the increase of the dimension of features, the cost of both storage and time increases dramatically. Thus a heuristic method is used in this paper to find approximate Markov blankets for the selected relevant features.

3.1 Approximate Markov Blanket

Koller and Sahami pointed out that if the training data was not enough, a big cardinal number of the Markov blanket will cause overfitting. So the cardinal number can be limited and a definition of approximate Markov blanket can be obtained based on *Corollary 1*.

Definition 2. Let M_i be the subset of *p* features in *S* which have the biggest MI with X_i. It is said that M_i is an approximate Markov blanket for X_i if

$$\frac{I(\{M_i, X_i\}, Y) - I(M_i, Y)}{I(M_i, Y)} < \alpha \quad . \tag{3}$$

where α is the redundant parameter. The larger α is, the more redundancy matters in feature selection and vice versa. The parameter *p* is the cardinal number of Markov blanket, and it can be altered according to the number of features.

When M_i is an approximate Markov blanket for X_i, the change rate of MI is limited to the range of smaller than α. When $\alpha \to 0$, $I(\{M_i, X_i\}, Y) \approx I(M_i, Y)$. Because $M_i \subset S$, so $I(\{S, X_i\}, Y) \approx I(S, Y)$, namely X_i is redundant for *S*.

In *Definition 2*, *S* is replaced by M_i to measure the redundancy between features, so that the size of feature subset and the computation cost can be reduced, and the accuracy can be increased comparing with methods considering single feature only.

3.2 Forward Feature Selection

We propose a feature selection algorithm based on approximate Markov blanket (FS_AMB) according to *Definition 2*. The algorithm adopts forward search scheme and *k*-NN MI estimation. The redundancy is measured by approximate Markov blanket. The cardinal number *p* of Markov blanket and the redundant parameter α are pre-specified.

The general steps of the forward feature selection algorithm can be described as follows:

(1). The first element of feature subset *S* is the feature which has the biggest relevancy (MI) with the output; (line 6)

(2). Select the next biggest relevant feature X_i as the candidate feature (line 8), and determine whether it is redundant or not based on the Markov blanket (line 9-19). If it is redundant, the subset *S* stays the same, namely *S*=*S* (line 15-16). Otherwise X_i is put into *S*, namely $S = S \cup X_i$ (line 17-19);

(3). Repeat step (2) until all the features are selected (line 7-20);

(4). Algorithm stops.

Algorithm1 (FS_AMB): Feature selection algorithm based on approximate Markov blanket

```
Input:  $F$ // the candidate feature set with $M$ features
        $Y$ // the output
        $p$// the cardinal number of Markov blanket
        $\alpha$ // the redundant parameter
Output: $S$ // the selected feature subset
Begin
1   Origin: $S = \{\}$
2   For $i = 1 : M$
3       Compute $I(X_i, Y)$ ;
4   End
5   $G = sort(I(X_i, Y), 'descend')$; // Sort features in descending order based on relevancy
6   $S = getFirstElement(G)$; //Put the most relevant feature into $S$
7   For $i = 2 : M$
8       $X_i = getNextElement(G)$ ;
9     If $p > |S|$
10        $M_i = S$ ;
11    Else
12        $S_{list} = sort(I(X_s, X_i), 'descend')$; // Sort features in descending order based on MI
13        $M_i = getFirstPElement(S_{list})$ ;// Get first $p$ relevant features
14    End
```

```
        If $\frac{I(\{M_i, X_i\}, Y) - I(M_i, Y)}{I(M_i, Y)} < \alpha$
            $S = S$; // If $M_i$ is an approximate Markov blanket, $X_i$ is a redundant feature. Remove $X_i$.
        Else
            $S = S \bigcup X_i$ ;// If $M_i$ is not approximate Markov blanket, $X_i$ is not a redundant feature. Select it into S.
        End
    End
End
```

4 Simulation Results

To testify the validity of the proposed algorithm, three data sets are used for simulation. They are artificial data set Friedman and Lorenz, and benchmark data set gas furnace. The performance is measured by root mean square error (RMSE) and normalized mean square error (NMSE).

4.1 Simulation Results for Friedman Data Set

The Friedman model is shown as equation (4), where $X_1, \dots, X_5$ are relevant features, $X_6, \dots, X_{10}$ are irrelevant features, and $X_{11} = 0.5 \times X_1$ is a redundant feature.

$$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + sigma(0,1) \quad . \tag{4}$$

This dataset can be used to examine whether the algorithm can select relevant features or not while there are both irrelevant and redundant features. The sample size is 500 and the simulation results are shown in table 1.

Table 1. Selected features with several selection methods for the Friedman data set

Selection method	Selected features
mRMR	X_4,X_2,X_1,X_5,X_6
NMIFS	X_4,X_2,X_1,X_5,X_6
FCBF	X_4,X_2,X_1,X_5,X_6
FS_AMB	X_4,X_2,X_1,X_3,X_5

Table 1 shows that mRMR, NMIF and FCBF select the same feature subset which consists of 4 relevant features and 1 irrelevant feature. However, the proposed method FS_AMB selects all the 5 relevant features.

4.2 Simulation Results for Lorenz Data Set

To further examine the validity of FS_AMB, the multivariate time series Lorenz is used for simulation. The Lorenz model is described by equation (5). When a=10, b=28, c=8/3, and the initial values are $x(0)=12, y(0)=2, z(0)=9$, equation (5) performs chaotic characteristics.

$$\begin{cases} \frac{dx}{dt}=a(-x+y) \ , \\ \frac{dy}{dt}=bx-y-xz \ , \\ \frac{dz}{dt}=xy-cz \ . \end{cases} \tag{5}$$

The fourth order Runge-Kutta is used to get the time series $x(t)$, $y(t)$, $z(t)$ whose step length is set as 0.02. Then the three dimensional time series are phase reconstructed, where the embedding dimension is 6 and the delay time is 8, 7 and 8 respectively. Thus we finally get 18 dimensional features.

$$X(t)=[x(t),x(t-8),...,x(t-5*8),y(t),y(t-7),...,y(t-5*7),z(t),...,z(t-5*8)] \ . \tag{6}$$

The output of single step forecasting are:

$$Y(t)=[x(t+1),y(t+1),z(t+1)] \ . \tag{7}$$

In this simulation, 1500 samples are used as training data and 500 samples are used as testing data. The proposed FS_AMB is used to select features and the forecasting model is then built with the selected features by general regression neural networks (GRNN).

The time series $x(t)$ is taken as an example to illustrate the performance of FS_AMB and Fig.1 and Fig.2 shows the predicted results with feature selection and without feature selection respectively.

The two figures show that the predicted error without feature selection is larger and the predicted output does not fit so well with the real output. On the other hand, the error curve tends to be steady and the predicted accuracy is well improved. Table 2 shows the selected features and predicted error of time series $x(t)$, $y(t)$, $z(t)$ with several selection methods.

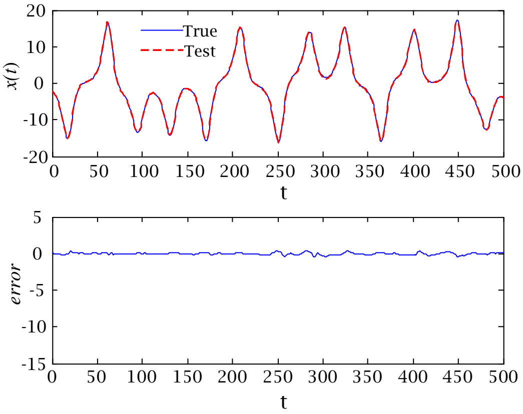

Fig. 1. Predicted output versus real output and error of GRNN model based on features selected by FS_AMB

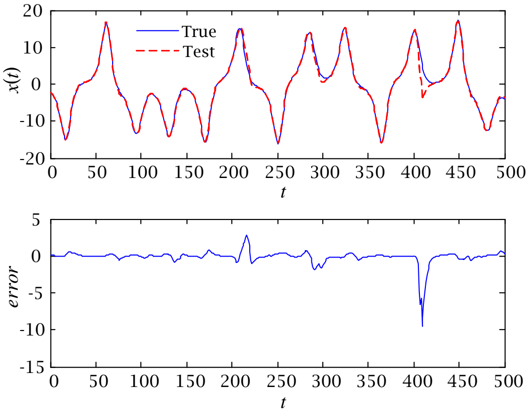

Fig. 2. Predicted output versus real output and error of GRNN model based on all the 18 dimensional features

Table 2. Selected features and predicted error with several selection methods for the Lorenz data set

	$x(t+1)$		$y(t+1)$		$z(t+1)$	
Selection Method	Selected Features	E_{RMSE}	Selected Features	E_{RMSE}	Selected Features	E_{RMSE}
mRMR	$x(t),y(t),z(t-16)$	0.1689	$y(t),x(t-8),z(t-8)$	0.2563	$z(t), x(t)$	0.2238
NMIFS	$x(t),y(t),z(t-16)$	0.1689	$y(t),x(t-8),z(t-8)$	0.2563	$z(t), x(t)$	0.2238
FCBF	$x(t),y(t-35),z(t-16)$	0.5302	$y(t),x(t-8)$	0.4002	$z(t), x(t)$	0.2238
FS_AMB	$x(t),y(t),x(t-8)$	0.1404	$y(t),x(t-8),x(t)$	0.2284	$z(t),z(t-8)$	0.2398

FS_AMB outperforms the other three feature selection methods in time series *x*(*t*) and *y*(*t*). In time series *z*(*t*), though FS_AMB does not perform best, the predicted error is comparable with the best results. In table 2, the user has to specify the desired number of features in the first three methods, while there is a stopping criterion in the rest two methods which can make the algorithm stop automatically. So there are only two features when predicting *y*(*t*) with FCBF. Another advantage of feature selection we can see from table 2 is that the network model is compact.

4.3 Simulation Results for Gas Furnace Data Set

In the gas furnace system described by Box and Jenkins, the input was the varied gas rate and the output was the CO_2 concentration in the outlet gas, forming two time series. The goal is to predict the CO_2 concentration of the output gas using the past values of both features. Ten candidate features are considered for building a predictive model of the gas furnace time series

$$u(t-6), u(t-5), \ldots, u(t-1), y(t-4), \ldots, y(t-1) \quad . \tag{8}$$

A GRNN model is trained with inputs selected by different methods. The size of the training data is set as 194, and the size of the testing data is set as 97. Table 3 shows the selected features and predicted error with several selection methods. The result of NMIFS is referred from [4]. To make the comparison fair, normalized mean square error is used in this section.

Table 3. Selected features and predicted error with several selection methods for the Gas Furnace data set

Selection Method	Number of Features	Selected Features	NMSE
mRMR	4	*y*(*t*-1),*u*(*t*-6),*u*(*t*-4), *y*(*t*-2)	0.042
NMIFS	4	*y*(*t*-1),*u*(*t*-4),*u*(*t*-5),*u*(*t*-6)	0.042
FCBF	3	*y*(*t*-1),*u*(*t*-6),*y*(*t*-4)	0.060
FS_AMB	4	*y*(*t*-1),*u*(*t*-6), *y*(*t*-2),*y*(*t*-3)	0.029

Among all the four selection methods, FS_AMB has the best performance, while FCBF has the worst performance. FCBF selects only three features which may cause the deficiency of information. So we can not say that the less the selected features there are the better the performance is.

5 Conclusions

To solve the feature selection problem in multivariate time series analysis, a novel forward feature selection method based on approximate Markov blanket is proposed. A new definition of approximate Markov blanket is given. To improve time efficiency, *k*-NN is utilized to estimate high dimensional MI. The MI between features and the output is used as the relevant criterion, and the approximate Markov blanket is used as the redundant criterion. The proposed FS_AMB method does not need to predifine the number of selected features. Simulation results show that it can not only select relevant features but also remove the redundant features. With this method compacter models can be built and better prediction performance can be achieved.

References

1. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
2. Battiti, R.: Using mutual information for selection features in supervised neural net learning. IEEE Trans. Neural Networks 5(4), 537–550 (1994)
3. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
4. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. IEEE Transactions on Neural Networks 20(2), 189–201 (2009)
5. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research (5), 1205–1224 (2004)
6. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proc. Int. Conf. on Machine Learning, pp. 284–292. Morgan Kaufmann, San Francisco (1996)
7. Kraskov, A., Stogbauer, H., Grassberger, P.: Estimating mutual information. Physical Review E 69, 66138 (2004)
8. Herrera, L.J., Rubio, G., Pomares, H., Paechter, B., Guillén, A., Rojas, I.: Strengthening the Forward Variable Selection Stopping Criterion. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009. LNCS, vol. 5769, pp. 215–224. Springer, Heidelberg (2009)