Partial Mutual Information Based Algorithm For Input Variable Selection

For time series forecasting

Ali Darudi [a], Shideh Rezaeifar [b], Mohammd Hossein Javidi Dasht Bayaz [c]
[a,c] Power Systems Studies and Restructuring laboratory
Dep. of electrical engineering
Ferdowsi University of Mashhad
Mashhad, Iran
ali.darudi@gmail.com

[b] Dep. of electrical engineering
Ferdowsi University of Mashhad
Mashhad, Iran

***Abstract*—In time series forecasting, it is a crucial step to identify proper set of variables as the inputs to the model. Many input variable selection (IVS) techniques fail to perform suitably due to inherent assumption of linearity or rich redundancy between variables. The motivation behind this research is to propose an input variable selection algorithm which not only can handle nonlinear problems and redundant data, but also is straightforward and easy-to-implement. In the field of information theory, partial mutual information is a reliable measure to evaluate linear/nonlinear dependency and redundancy among variables. In this paper, we propose an IVS algorithm based on partial mutual information. The algorithm is tested on three time series with known dependence attributes. Results confirm credibility of the proposed method to capture linear/non-linear dependence and redundancy between variables.**

Keywords- input variable selection; partial mutual information; time series forecasting; information theory.

I. INTRODUCTION

Input variable selection (IVS) is the procedure of selecting a proper subset of variables from all potential inputs to a model. In prediction tasks, suitable input variables not only have maximum dependency with prediction variable (target) but also demonstrate minimum redundancy among themselves[1].

Forecasting problems has attracted much attention in electricity markets in order to predict market signals including demand, market clearing price and, recently, wind generation. Proper input variable selection is a crucial step in any time series forecasting procedures[2]; because data driven techniques are greatly sensitive to input variables fed to the model. Both excessive and deficient numbers of inputs degrades prediction performance of the model. Excessive number of inputs might have the following consequences: (i) irrelevant inputs have negative impacts on the learning procedure of the model [3] (ii) computational burden and complexity increases without improvements of model accuracy (iii) the true drivers of the modeled system become difficult to recognize. On the other hand, if variables relevant to target are ignored, forecasting model will be unable to properly distinguish input-output relationships [4].

Input variable selection methods could be categorized into model-based and filter (model-free) techniques (Fig. 1). Model-based approaches search the space of variable subsets, using the training/validation accuracy of a particular forecasting model. Potential proper variables are fed to the model; then, utility of the variables is measured based on forecasting performance of the model [5]. In contrast, filter methods do not depend on any pre-existing model; they separate model training and input variable selection tasks and incorporate statistical analysis methods to measure significance of input variables [6]. Recently, filter methods have gained popularity due to their independency to any specific model and higher computational speed in comparison to model-based methods[5].

Filter methods are classified into correlation (linear) and information theoretic (non-linear) approaches. As the name indicates, linear methods cannot capture non-linear dependencies between data; therefore, effective variables might be omitted from the selected input set. Besides, they are sensitive to noise and data transformation[7], which might be applied in pre-processing steps of forecasting procedure[2].

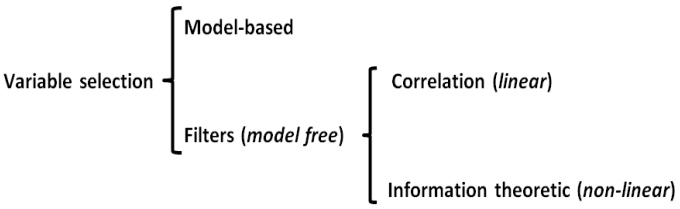

Fig.1 Taxonomy of input variable selection algorithms

978-1-4799-2803-3/13/$31.00 ©2013 IEEE

Recently, techniques based on information theoretic (non-linear) approaches have been applied in several IVS algorithms. They adapt properly with popular non-linear forecasting techniques such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

In information theory field, *mutual information* (MI) is widely used to measure and analyze information. Mutual information between two variables indicates the predictability level of a variable if another one is known [8]. In the literature, several researches have used mutual information as a measure of dependency for non-linear problems [1][9][10]. However, MI does not take into account possible redundancies among selected input variables [4]. As an example, suppose variables X and Z are highly relevant to the target value and also they are dependent on each other (say X=2Z). They are both selected as input variables, because they have high values of MI with the target value. However, Z should not belong to selected set as it is a redundant variable, described totally by X.

In order to overcome this limitation of MI algorithms, partial mutual information (PMI) was proposed[11]. PMI(Y,X|Z) is analogous to the partial correlation coefficient and quantifies the dependence of Y (target value) on input variable X that is not accounted for by the input variable Z [2]. In other words, PMI measures additional information that a new variable provide for a pre-existing prediction model[12]; therefore, PMI is potentially a proper tool for distinguishing redundant variables.

In this paper, we propose a model free IVS algorithm called *PMI based max relevance-min redundancy*. We acquire PMI measurements to deal with non-linear and redundant variables. Moreover, unlike other algorithms, dependencies between selected and non-selected variables are taken into account.

This paper is organized as follows: In section II we first present the background on MI and PMI concepts and review conventional minimal redundancy maximal relevance (mRMR) criterion. Section III presents a new PMI based input variable selection algorithm. Section IV gives experimental results on three data sets. Finally, paper is concluded in section V.

II. BACKGROUND

A. Mutual information

Mutual information between two random variables x and y is denoted as:

$$I(x;y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy \qquad (1)$$

Where $f(x)$ and $f(y)$ are the probability density functions of X and Y and $f(x,y)$ is the joint probability density function of X and Y. In the case of no dependency between two variables, joint probability density $f(x,y)$ would be equal to the product of probability densities, so MI equals to zero.

Mutual information measures dependencies between variables without any assumption concerning the linearity of dependency[8]. In fact, MI is considered as the information each variable describes about the other one[6].

B. mRMR

The ultimate aim of input variable selection is to recognize input set, which are most dependent on the output. This is ideally achieved by Max-Dependency according to following relationship:

$$Max\ D(S,y), \qquad D = I\ (\{x_i, i = 1, \dots m\}, y) \quad (2)$$

Where S is selected variable set with m variables and y is the target value.

However, calculation of D involves computation of multivariate density $p(x_1 ,..,x_m)$ and $p(x_1 ,..,\ x_m\ ,y)$ which introduces some difficulties including: i) large number of samples are required which might be unavailable ii) in order to estimate multivariate density function, it is required to compute inverses of high dimensional matrices which is burdensome[13][14].

To deal with these limitations, especially in the case of large number of variables, MI based methods usually tend to use bivariate statistics [6]. For instance, an alternative approximation for Max-Dependency based on maximal relevance (MR) could be used which is defined as:

$$Max\ D(S,y),\ \ D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;y) \qquad (3)$$

Selecting input variables based on this criterion may not yield optimum input variable set. Reference [15] emphasizes that there is a distinction between relevance to target and usefulness of a variable. Redundancy between variables may make some variables useless for forecasting purposes[6].

Max relevance algorithm lacks a procedure for discriminating redundancy; in order to avoid this issue, an auxiliary minimal redundancy (mR) condition may be considered as follow:

$$Min\ R(S)\ ,\ \ R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j) \qquad (4)$$

In order to exploit advantages of both above criteria, minimal-redundancy maximal relevance (mRMR) [13] is defined as:

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

Consequently, input variables are selected based on both relevance with respect to target value and independence with respect to other selected variables. Intuitively, mRMR yields better performance in comparison with mR and MR alone[6][16].

In practice, mainly incremental search strategies are applied to find the suitable set of variables. Incremental search methods selects one individual variable at each step and add to the selected set; the most significant variable is selected and then the procedure continues iteratively according to the fulfillment of the criteria[6].

C. Partial Mutal Information

MI cannot deal with redundancy among variables directly. Partial mutual information (also known as conditional mutual information) was proposed in [11] to rectify this problem . As a definition, partial mutual information is the amount of information shared between two variables considering pre-existing variables. The PMI between variable x and target y with respect to selected set of w is defined as[4]:

$$PMI(x';y') = \iint f(x',y') log \frac{f(x',y')}{f(x')f(y')}\, dx'\, dy' \quad (6)$$

Where $x' = x - E[x|z]$; $y' = y - E[y|z]$

Operator $E[.]$ refers to the expectation of variable. The variables x' and y' are residuals of variables x and y regarding selected set z. The discrete version of partial mutual information criterion is given by[4]:

$$PMI = \frac{1}{N}\sum_{i=1}^{N} log \frac{f(x',y')}{f(x')f(y')} \quad (7)$$

Interpretation of different values for MI and PMI are summarized in Table I.

III. Proposed Method

mRMR algorithm has a few deficiencies including: i) Min-redundancy criterion has been defined loosely as: to minimize relevance between selected variables. However, to be more precise, the relevance between selected variables with respect to the target is the critical factor for detecting redundancy. ii) Relationships between selected variables and non-selected variables have not been explicitly considered. For instance, if a non-selected variable has a high MI value, there must be at least one variable in the selected set that is highly relevant to it. No criterion exists in mRMR to ensure such features.

TABLE I. Interpretation of different values for mutual information (MI) and partial mutual information (PMI)

	High value	**Low value**
PMI(y,x|z)	X is independent of z with regard to forecasting target y	Z is a well-suited representative of x with regard to forecasting target y
MI(x,y)	X and Y is highly correlated	X and Y is almost independent

To overcome these shortcomings, we propose a method that utilizes partial mutual information (PMI) to modify redundancy criteria and captures dependencies between selected and non-selected variables. Following paragraphs will firstly define and elaborate three desired characteristics of a proper input variable set. Then they are formulated based on MI and PMI. These characteristics and formulations are defined such that only bivariate MI and PMI calculations are required; thus the whole algorithm is robust, easy-to-implement and time efficient.

A proper set of selected input variables own the following features:

1- *Max relevance of the selected inputs to target value:* selected inputs should have the largest relevance to the target variable; which is basically the same as maximal relevance criterion. Therefore sum value of all mutual information values between selected inputs and target variable should be maximized:

$$Max\ D(S,y),\ \ D = \frac{1}{|S|^2}\sum_{x_i \in S} I(x_i;y) \quad (8)$$

Where S denotes the selected set.

2- *Max independent relevance of selected variables to the target value:* selected inputs should show minimal redundancy with each other. In partial mutual information terms, it means that PMI between each two selected variables and target value should indicate high values which ensure selected variables are not only relevant to the target value, but also independent from each other (without redundancy). In other words, when PMI between two selected variables and target value is low (close to zero), one of those selected variables could be removed without having any negative effects on the utility of the selected input set.

We propose optimization problem in (9) to represent this criterion:

$$Max\ MS(S,y),$$
$$MS = \frac{1}{|S|^2}\sum_{x_j,x_i \in S} PMI\ (y, x_i|x_j) \quad (9)$$

3- *Min independent relevance of non-selected variables to target value:* non-selected variables are candidates that do not belong to selected input set. Any non-selected variable should have at least one of these two features: (i) it has insignificant relevance to target value (First criterion (9) ensures this feature is taken into account in our method). (ii) If it is relevant to target value, there is at least one variable in the selected input set which is highly relevant to the not-selected; hence it is unnecessary that selected set includes this non-selected variable. As mentioned earlier, low values of PMI(Y, X|Z) indicates that X is highly relevant to Z while trying to forecast target value Y. Therefore in PMI terms, for each non-selected variable, the PMIs between the non-selected variable and at least one of the selected

variables with respect to target value should indicate low values.

Fig. 2 exemplifies a proper set of selected variables with respect to this feature. Although variable X_{u2} has a high MI value with respect to the target value, it belongs to non-selected set; because low value for PMI(Y, $X_{u2}|X_{s1}$) ensures that X_{s1} is a well-suited variable representing X_{u2}; therefore X_{u2} is removed from selected set, correctly. However, being a careful observer, one might notice that PMI(Y, $X_{u2}|X_{s2}$) has a high value, but it does not have a negative impact on appropriateness of the selected input variables; because only one variable (X_{s2}) is sufficient to represent X_{u2} in the selected input set. In other words, only the minimum value of PMI between the non-selected value and each of the selected values is important. Other members of non-selected set should also demonstrate the same features.

We define the formula (10) to approximate above-mentioned criterion:

$$Minimize\ MU(S,U,y),$$
$$MU = \frac{1}{|U|^2}\sum_{x_j \in U} \min\Big(\ PMI\ (y\ , x_j|x_i)\Big)\ \forall\ x_i \in S \quad (10)$$

Where U is the subset of not-selected variables. The "*min*" operator inside the objective function ensures that only the minimum value of PMI between not-selected value and each of the selected values are considered. Sigma also indicates that the criterion should be studied for all members of not-selected set.

We call the criterion combining all three objective functions above *PMI based max relevance-min redundancy* (PMI-mRMR) which could be stated in the simplest form as the following optimization problem:

$$Minimize\ \ MU - MS - D \quad (11)$$

The optimization problem may be modified or redefined to satisfy arbitrary additional constraints; for instance, size of selected set (number of selected input variables) may be specified as a constraint to the problem.

In order to solve the optimization problem, we perform a global search, in opposition to conventional incremental search. Incremental search strategies commence at a location and move through the searching space by adding one variable at a time. Therefore, they may achieve a locally optimal solution and terminate prematurely. Optimality degree of the solution highly depends on the amount of search space that is explored. Accordingly, global methods yield to better solutions because they consider more possible combinations.

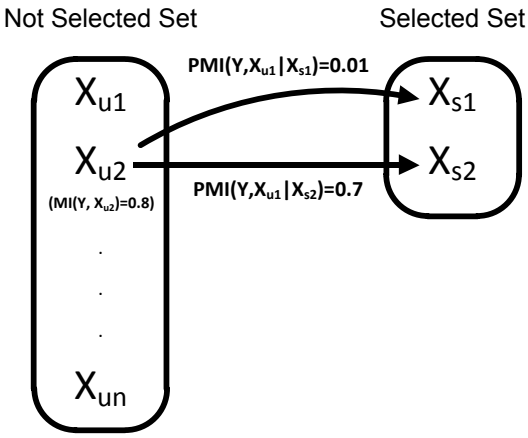

Fig.2 Suitable sets of selected and non-selected input variables with respect to X_{u2}. Although X_{u2} is highly relevant to target value (MI(Y,X_{u2}) is high), low value of PMI(Y,$X_{u2}|X_{s1}$) ensures that X_{s1} is a proper representative for X_{u2} in the selected set.

IV. Case Study

D. Test problems

In order to test performance of newly proposed input variable selection algorithms, they are fed several benchmark synthetic data. Use of synthetic data, in comparison to real-world data, is more useful because true effective variable sets are known beforehand; thus algorithms are evaluated more accurately [2].

The proposed algorithm is tested on three data sets generated by both linear and non-linear models. These models were also used in some other papers to evaluate credibility of proposed algorithms.

i) AR9- Linear Autoregressive time-series, order 9

$$x_t = 0.3x_{t-1} - 0.6x_{t-4} - 0.5x_{t-9} + e_t \quad (12)$$

Where e_t is a Gaussian random noise with a zero mean and unit standard deviation.

ii) TAR2-Threshold Autoregressive time-series, order2

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} + 0.1e_t & if\ x_{t-6} \le 0 \\ 0.8x_{t-10} + 0.1e_t & if\ x_{t-6} > 0 \end{cases} \quad (13)$$

iii) Non-linear system

$$y = (x_2)^2 + \cos(x_6) + 0.3\sin(x_9) \quad (14)$$

E. Data

For the first two time series models, 520 data points were generated; we discarded first 20 points to decrease the effect of an arbitrary initialization. Candidate set consists of the first 15 lags of X. For the non-linear model, 15 standard Gaussian random variables were generated (x_1 to x_{15}), each 500 data points long. Target value (y) was generated using x_2, x_6 and x_9; and x_1 to x_{15} were chosen as candidate set of potential inputs.

F. Results

The proposed method, *PMI based max relevance-min redundancy* (PMI-mRMR), was evaluated for two scenarios where true numbers of effective input variables: i) were given to the optimization problem as a constraint. ii) were not specified. True numbers of effective variables for test problems are respectively: 3, 2 and 3. The results are shown in table 2.

We used Genetic Algorithm (G.A) to solve optimization problem of the algorithm. However, we observed that the algorithm is such highly time efficient that even conducting an exhaustive search for the optimum solution is reasonable for the test functions evaluated. For instance, for the first scenario, it took only 0.18 seconds, on average, to execute exhaustive search for each of the models; all computations are performed on a Pentium IV with 3 GB of RAM at 2GHz. All MI and PMI measurements were done using *Feast* toolbox in MATLAB provided by [5] available at [17].

Proposed algorithm works properly if number of selected variables is given (first row in table 2); for TAR2 and Non-linear models, correct variables are selected. Although, in the case of AR9 model, one of the variables is identified incorrect (lag 11 instead of 1), further investigations indicated that second best answer of the optimization problem (11) is actually the expected values (lags 1, 4 and 11). Therefore, for forecasting applications, it is recommended to store first few best answers provided by the algorithm and test their predictability performance separately.

On the other hand, when number of effective variables is not already specified (second row in table II), although no irrelevant or redundant variables are selected, the algorithm tends to under-estimate size of the input variable set; one effective variable is ignored for each of the test problems.

In comparison with other algorithms, results confirm that the PMI based algorithm performs reasonably well with high computational efficiency. All other algorithms are unable to identify correct input variable set; they choose several redundant variables due to their inherent deficiencies.

TABLE II. RESULTS OF INPUT SELECTION ALGORITHMS ON TEST PROBLEMS

Algorithms	**AR9 (1,4,9)**	**TAR2 (6,10)**	**Non-linear system (X2,X6,X9)**
PMI-mRMR Scenario I	Lags 4 9 11[a]	Lags 6 10	X_2 X_6 X_9
PMI-mRMR Scenario II	Lags 4 9	Lag 10	X_2 X_6
Correlation [18]	Lag 4,13 and others	Lag 10 and others	X_2 and others
Trial- Method[18]	Lag 4,9 and others	Lag 10 and others	X_2 X_6
D-value [18]	Lag 4 and others	Lag 10 and others	X_2, X_6 and others
Supervised- SOM[3]	Lag 1,3,5, 7,8,9	Lag 3,4,5, 6	X_2 X_{14}

a. second best answer is the correct value: 1, 4, 11

V. Conclusion

Selecting proper Input variable is a crucial task in time series forecasting. The motivation behind this research was to formulate a simple, straightforward and easy-to-implement, yet effective, input variable selection algorithm. Therefore, a criterion based on partial mutual information was proposed and tested on benchmark time series with known dependence attributes. Results confirm credibility of the proposed method to capture linear and non-linear dependence between variables. Moreover, it is immune to redundancies between variables. Since the proposed method is based on only bivariate MI and PMI measurements, it is robust and computationally efficient. As a result, the algorithm potentially suits well for real-world problems with large and high dimensional data sets.

References

[1] R. K. Parviz, M. Nasser, and M. R. J. Motlagh, "Mutual Information Based Input Variable Selection Algorithm and Wavelet Neural Network for Time Series Prediction," ICANN, pp. 798–807, 2008.

[2] R. J. May, H. R. Maier, G. C. Dandy, and T. M. K. G. Fernando, "Non-linear variable selection for artificial neural networks using partial mutual information," Environmental Modelling & Software, vol. 23, no. 10–11, pp. 1312–1326, Oct. 2008.

[3] G. J. Bowden, G. C. Dandy, and H. R. Maier, "Input determination for neural network models in water resources applications. Part 1—background and methodology," Journal of Hydrology, vol. 301, no. 1–4, pp. 75–92, Jan. 2005.

[4] T. M. K. G. Fernando, H. R. Maier, and G. C. Dandy, "Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach," Journal of Hydrology, vol. 367, no. 3–4, pp. 165–176, Apr. 2009.

[5] G. Brown and M. Luj, "Conditional Likelihood Maximisation : A Unifying Framework for Information Theoretic Feature Selection," Journal of Machine Learning Research, vol. 13, pp. 27–66, 2012.

[6] R. May, G. Dandy, and H. Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks," in Artificial Neural Networks - Methodological Advances and Biomedical Applications, K. Suzuki, Ed. InTech, 2011, p. 362.

[7] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE transactionson on neural networks, vol. 5, no. 4, pp. 537–550, 1994.

[8] M. I. Hejazi and X. Cai, "Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm," Advances in Water Resources, vol. 32, no. 4, pp. 582–593, Apr. 2009.

[9] A. Papana, D. Kugiumtzis, and C. Sciences, "Evaluation of Mutual Information Estimators for Time Series," International Journal of Bifurcation and Chaos, Applied Sciences and Engineering, pp. 1–47, 2009.

[10] M. N. Maralloo, A. R. Koushki, C. Lucas, and M. M. Pedram, "Mutual Information Based Input Selection in Neuro-Fuzzy Modeling for Long Term Load Forecasting,"Computer Science & Information Technology (CSIT 2009), Yerevan, Armenia, 28 Sept.-2 Oct., 2009, pp. 209-213.

[11] a. Sharma, "Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 — A strategy for system predictor identification," Journal of Hydrology, vol. 239, no. 1–4, pp. 232–239, Dec. 2000.

[12] I. Luna and R. Ballini, "Top-down strategies based on adaptive fuzzy rule-based systems for daily time series forecasting," International Journal of Forecasting, vol. 27, no. 3, pp. 708–724, Jul. 2011.

[13] H. Peng, "Feature Selection Based on Mutual Information:Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE

transactions on pattern analysisand machine intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

[14] F. Rossi, a. Lendasse, D. François, V. Wertz, and M. Verleysen, "Mutual information for the selection of relevant variables in spectrometric nonlinear modelling," Chemometrics and Intelligent Laboratory Systems, vol. 80, no. 2, pp. 215–226, Feb. 2006.

[15] R. Kohavi and H. John, "Artificial Intelligence Wrappers for feature subset selection," Artificial Intelligence, vol. 97, no. 97, pp. 273–324, 1997.

[16] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," Journal of Bioinformatics and Computational Biology, vol. 3, no. 2, pp. 185–205, 2005.

[17] feast toolbox available at: www.cs.man.ac.uk/~gbrown/fstoolbox/

[18] C. Yuan, X. Zhang, and S. Xu, "Partial mutual information for input selection of time series prediction," 2011 Chinese Control and Decision Conference (CCDC), no. 1, pp. 2010–2014, May 2011.