Application of artificial intelligence techniques in building economic-financial forecast models on high dimensional data sets

Do Van Thanh

Hanoi, May 5 - 2018

Goals and contents

1. Goals

1) Overview of the application of artificial intelligence techniques in the modeling of economic-financial forecasts on high dimensional data sets.
2) Challenges and proposed solutions
3) Application example

2. Contents

1) Scope of presentation
2) Arised problem
3) Brief history of the problem
4) Methodology for solving
5) Dimensionality reduction: content, challenge and our proposals
6) Statistical & AI techniques: Advantages and Disadvantage; our Proposals
7) Examples.

Define the scope of presentation

- What is forecast?
- Three English terms mean “Forecast”
  - ✓ Prediction: Just based on past rules
  - ✓ Forecast: based on past rules and taking into account the abnormalities of the future
  - ✓ Foresight: Rules formed in the past are not used much to forecast the future
- “Forecast” is most important, given the most attention
- There are 3 types of forecasting?
  - ✓ Forecast the time of occurrence of the event
  - ✓ Forecast the impact of the event
  - ✓ Forecasting time series : is the most popular and important, especially in the financial-economic field

➢ **This presentation focuses on forecasting time series in the field of economics - finance**

Arised Problem

- Problem:
  - ✓ Let Y denote the target variable, and X_1, X_2, ..., X_n denote original candidate variables; Y and X_i, are m-dimensional vectors, the numbers m and n are very big.
  - ✓ Question: How to predict the Y variable according to the X_i (i=1, 2, ..., n) variables?
- The problem is considered in three cases:
  1. The original and target variables are continuous or categorical
  2. The variables have functional valuses
  3. The variables have symbolic valuses

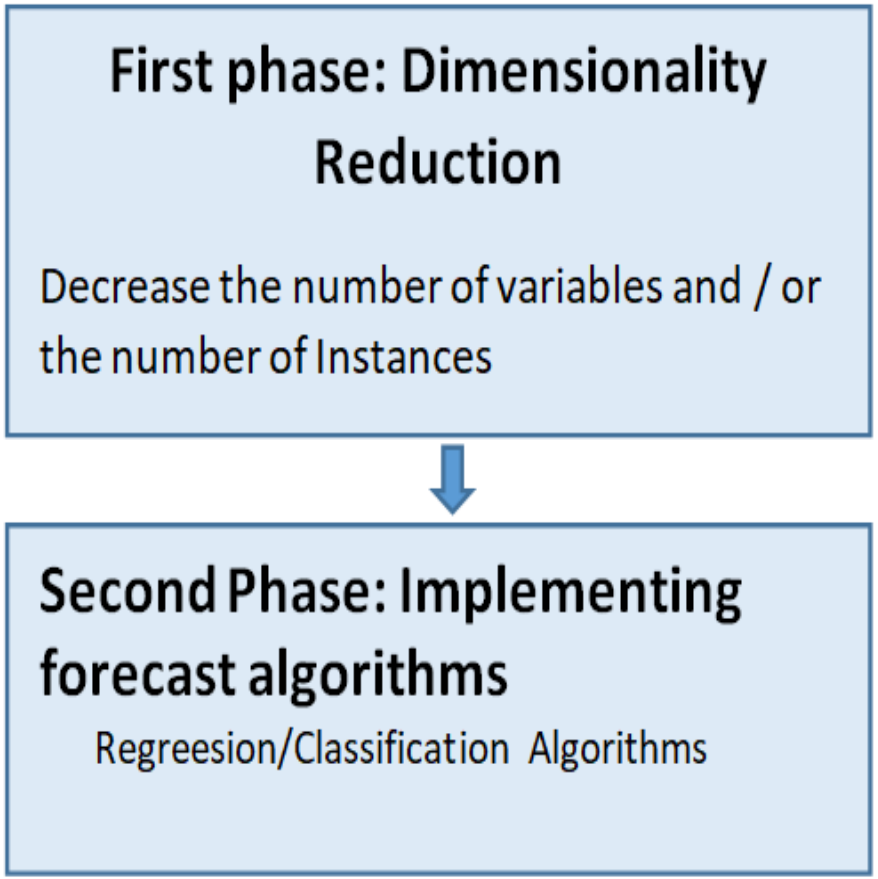
➢ This presentation focuses on the first case

Brief history of the problem

- When n, m is not big, in which is n, was considered and applications in the approximately 100 years recent
- Methedology: Using statistical techniques, especially techniques of multivariable regression and logistic regression;
- Multivariable regression includes many different regression methods
- When n, m is very big (especially n)
  - ✓In order to do regression: Economists generally choose a few root variables that are highly correlated with the target variable and are related to the target variable according to economic theories.
  - ✓Cons: The selected variable may be not most fit, many important information may be omitted.
  - ✓Note: the more information the model contains, the higher the forecast accuracy
- The curve "high dimensional data" has existed for quite long time
- Dimensionality Reduction (in the mid-1990s) to overcome this curse

Methodology for solving

- Reducing the number of variables is the most important of dimensionality reduction;
- Techniques of Dimensionality Reduction include: Variable subset selection and Variable Transformation.
- The techniques used to build the forecasting model include: statistical techniques and artificial intelligence ones



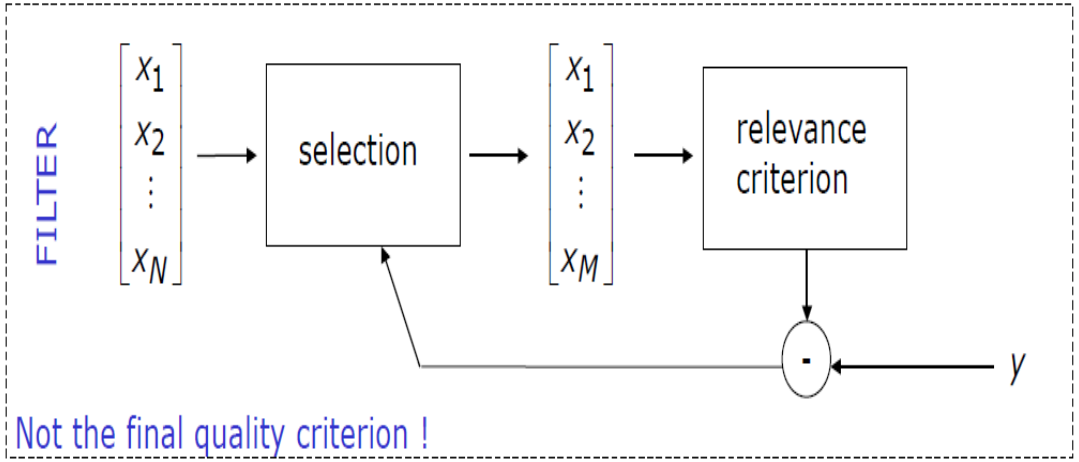
Dimensionality Reduction

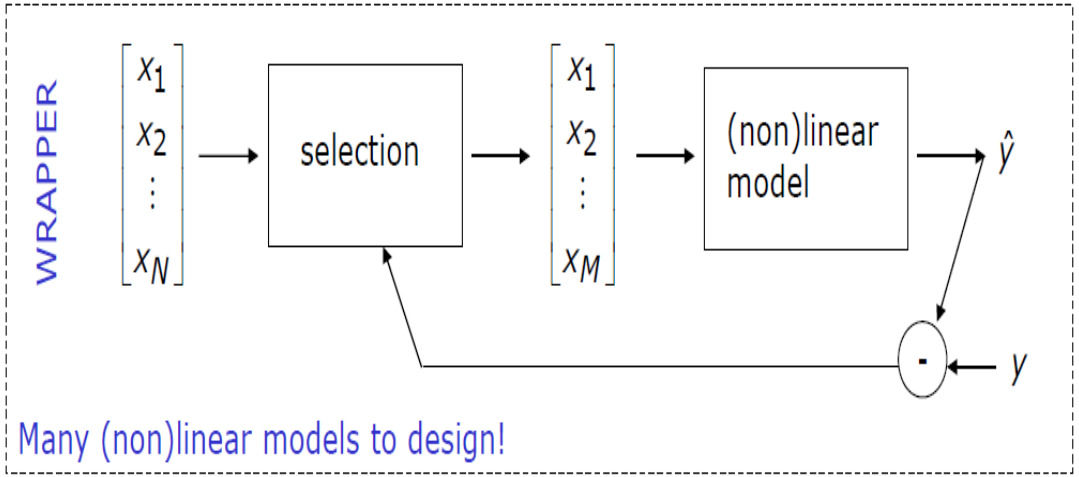
- Selection of subset of variables is done in 3 approaches: filter, embeded, wrapper.
- Transformation of variables: genetic algorithms, hill climbing algorithms, PCA, …
- In practice: dimensionality reduction is done by combining several techniques in the above approaches.

$$\begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \\ \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_N \end{bmatrix} \right)$$

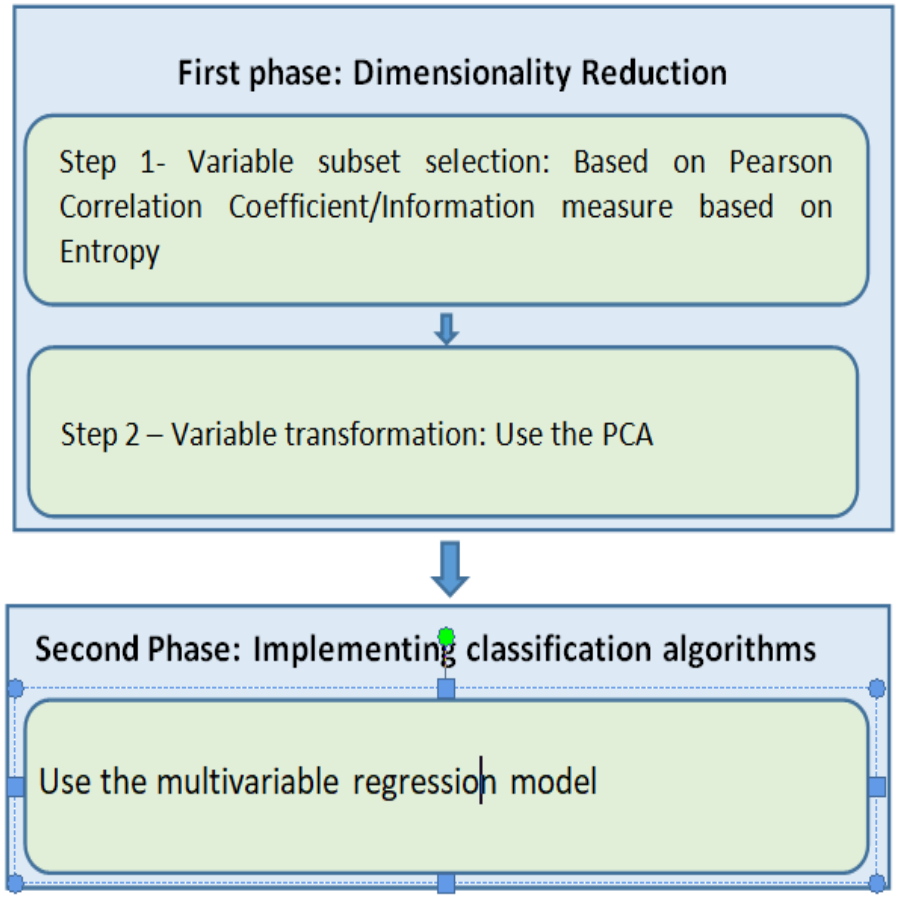
Filter vs. wrapper

- Filter:
  - ✓ For continuous variables: use the correlation measure;
  - ✓For categorical variables: use the mutual information measure, …
- Wapper
  - ✓Hill climbing algorithms, decision trees, genetic algorithms, artificial neural networks, …





Dimensionality reduction in the real world



- **Why** ?

✓ There are some correlation measures, the Pearson correlation coefficient is still the most effective and it is used to evaluate the fit of econometric models.

✓The pearson correlation coefficient measures the linear correlation between two variables, while the mutual correlation measure for nonlinear correlation

✓Experiments show that PCA are still the most effective technique for dimensionality reduction of real data set

Evaluate the efficient of PCA

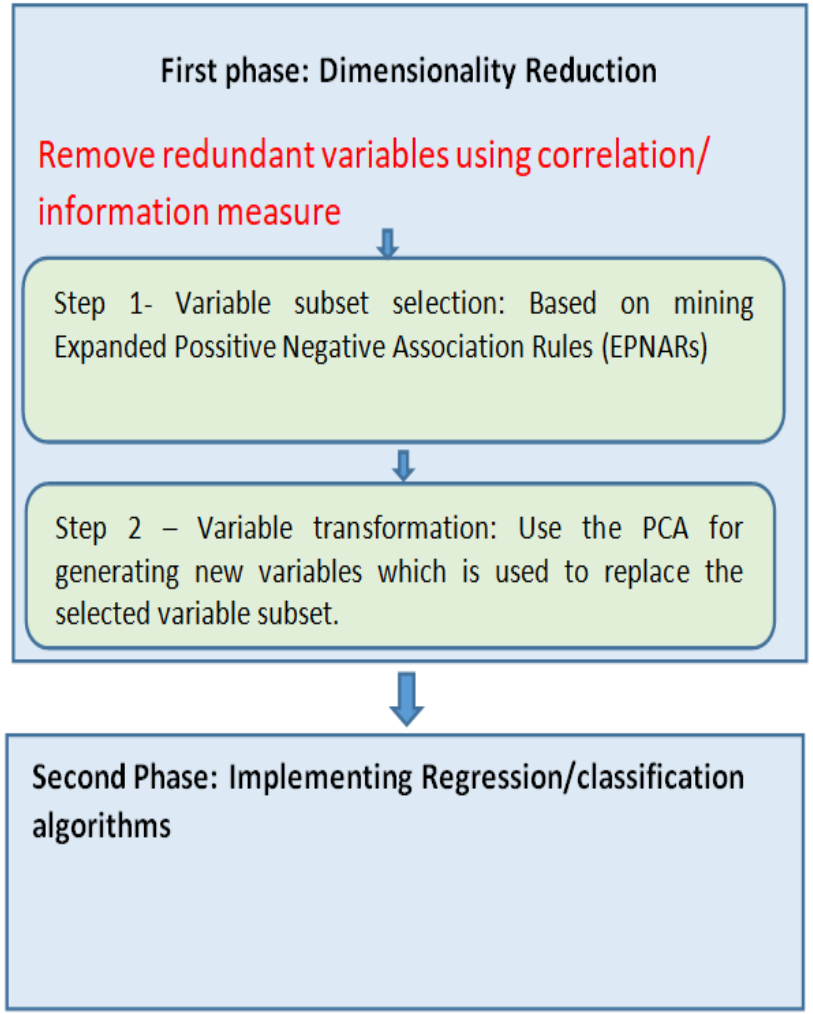
- Van Der Maaten et al (2009 - Journal of Machine Learning Research) compared 12 leading nonlinear techniques of dimensionality reduction with PCA: *Multidimensional Scaling, Isomap, Maximum Variance Unfolding, kernel PCA, Diffusion Maps, Multilayer Autoencoders, Locally Linear Embedding, Laplacian Eigenmaps, Hessian LLE, Local Tangent Space Analysis, Locally Linear Coordination, and Manifold Charting,*
- Conclusions*:* The 12 technical work well on the created data sets but on data sets of real world no any technique is more effect PCA.

- Note: PCA is efficient when n is not too big and that the data points need to be approximately a hyperplane or manifold.
- KPCA is similar to KSVM in order to overcome the limitations of PCA, but:
  - ✓Difficult to apply in fact,
  - ✓Big loss cost to choose an appropriate kernel function.

Limitation of Pearson Correlation Coeficient and Mutual Information Measure

- Limitations
  - some original variables, if considered individually, are highly correlated with the target variable, but this is not always the case for the set of these variables
- Overcome
  - Current popular way: Select variables that are highly correlated with the target variable and remove redundant variables
  - Achilles: Variable has low correlation coefficient with the target variable can still be very useful for forecasting.
- Question: Is there a way to overcome this phenomenon?

Our Proposal



Definition 1: Extended positive negative association rule (EPNAR) is an association rule of the form $r = \neg A \cup B \Rightarrow \neg C \cup D$, where at least one of the two itemsets $\neg A$ or B as well as one of two itemsets $\neg C$ or D must be non-empty and the itemsets $\neg A$, B, $\neg C$ and D do not intersect.

Can see that if $\neg A = \phi$, $\neg C = \phi$, $B \neq \phi$ and $D \neq \phi$ then r is a positive association rule, and if ($\neg A \neq \phi$, $\neg C = \phi$, $B = \phi$ and $D \neq \phi$) or ($\neg A \neq \phi$, $\neg C \neq \phi$, $B = \phi$ and $D = \phi$) or ($\neg A = \phi$, $\neg C \neq \phi$, $B \neq \phi$ and $D = \phi$) then r is a negative association rule. But if at least 3 of the 4 itemsets $\neg A$, $\neg C$, B and D are not empty then r is not a positive as well a negative association rule.

Application of dimensionality reduction

In many areas

- ✓ Image recognition
- ✓ Speech recognition
- ✓ Analysis of environmental data, bioinformatic data,
- ✓ Analysis and forecast of finalcial-economic data, ...

Forecast/classification algorithms are used primarily

1. Multivariate regression
2. Time series analysis: VAR, VECM, ...
3. Bayesian classìier
4. Logistic regression
5. Decision tree
6. Neutron network
7. Genetic algorithm
8. Association Rules
9. SVM, KNN, K-mean,
10. And many other classification algorithms

They are divided into two groups: Statistical Techniques and Artificial Intelligence Techniques

Statistical Techniques vs. Artificial Intelligence ones

Statistical Techniques (regression)

- Main advantages
  - ✓Forecasting model is clear, delicate; have been interested in a very long time, convenient to forecast continuous variables; forecast Accuracy is high; can determine behaviors, be used for Prediction, Forecast, Foresigh; can evaluate impacts of shocks and policies, ...
- Main disadvantages
  - ✓ Not fully automated; must perform many statistical tests, including testing input variables?
  - ✓ Failed to execute on large data sets, ...

Artificial Intelligence Techniques

- Main advantages
  - ✓Can be done on large data sets, do not need to perform statistical tests, can be fully automated, convenient to forecast classification;
- Main disadvantages
  - ✓ Forecasting model is hidden, can not be evaluate for impacts of shocks & policies
  - ✓Theoretically it is possible to forecast with high precision, but to achieve it needs so much costs;
  - ✓Mostly used for prediction.

Challenges for artificial intelligence techniques

- Economic - financial data are mainly numeric, Artificial Intelligence algorithms are very weak on point value prediction
- Behavioral research and impact assessment of economic shocks and policies are very important in the economic - financial field.
- Theoretically, it is great, but the fact is not as expected and to improve predictability: it expenses many time, memory and other costs;

Our proposals

- Combining both statistical forecasting techniques and artificial intelligence ones to improve the forecast accuracy on high dimensional data sets in the field of economics – finance
- Specifically, after creating a new variable subset to replace the set of original variables:
  - ✓Use regression techniques to forecast point values and interval values;
  - ✓Divide the forecast intervals into a smaller ones. Use artificial intelligence techniques to predict the class of input data tuples
- Note: the narrower the forecast, but the higher the accuracy, the better the forecast quality

Combine quantitative forecasts with sentiment analysis

- Forecast by quantitative models is assuming that the future takes place the close as the past and present. But the reality is not so.
- To improve forecast accuracy :
  - The current approach of economists: Combining an use of quantitative models and judgmental method;
  - Multidisciplinary approach: Combining an use of quantitative models and sentiment Analysis.
  - Our other approach: Simulation of risk levels of forecast. Method: using logistic regression technique.

Example

- **Problem**: forecasting the price of FPT stock according to 51 economic and financial variables that are potential to impact on FPT stock price under economic theories.
- **Data**: Mostly by date: from January 3, 2012 to May 31, 2017; Monthly data: from 1/2012 to 5/2017. Transfer data by date into data by month.
- **Characteristics**:
  - ✓This is the most difficult forecast in the field of economics – finance
  - ✓Before 1978, economics believed that: can not forecast the price.
  - ✓After 1978, believe that the price can be forecasted if the market run not efficiently?
  - ✓How to know: by using the model GARCH (p,q) (1982, by Engle – Nobel, 2003).
- **In case of FPT stock price**:
  - ✓Market runs inefficiency: the reaction of strategic investors of this stock is not timely; Market reaction is slow because of inertia.

Methedology

- **Dimentionality Reduction**

  Time 1: selected 28 in the 51 variables;

  - Time 2 (PCA): create 5 new variable from the 28 selected variables - retain 90.9% of information of the 51 of the original variables

- **Theoretical model of FPT stock price forecasting**: Using Autoregressive Lagged Model (ADL) and GARCH model (to forecast the residual variance of the model ADL as well as to evaluate efficient of the market of FPT stock.

Mô hình lý thuyết dự báo giá cố phiếu FPT

$$Y = c + \sum_{i=0}^{r1} a_{1i} X_1(-i) + \sum_{i=0}^{r2} a_{2i} X_2(-i) + \ldots + \sum_{i=0}^{rk} a_{ki} X_k(-i) + \sum_{q=1}^{r} b_q Y(-q) + u(t) \quad (1)$$

$$H = \propto + \sum_{i=1}^{p} a_i H(-i) + \sum_{i=1}^{q} b_i \,.\, u(-i)^2 + \sum_{i=0}^{r1} c_{1i.} X_{k+1}(-i) + \sum_{i=0}^{r2} c_{2i.} X_{k+2}(-i) + \cdots + \sum_{i=0}^{r(g-k)} c_{(g-k)i.} X_g(-i) + \epsilon(t) (2)$$

ở đây: H(t) là phương sai của u(t);

$X_1, X_2, \ldots, X_k$ (k=5) là các biến mới (thành phần chính) thay thế tập biến gốc ban đầu;

$X_{k+1}, \ldots Xg$: các biến được sử dụng để đánh giá hiệu quả của thị trường cổ phiếu FPT

X(-i) là biến X trễ i tháng.

Các hệ số của của PT (1) và (2) được ước lượng bằng phương pháp OLS.

Test forecast results

Training data set: From 1/2012 to 12/2016

Testing data set: From 1/2017 – 5/2017

Quan sát	Actual	Forecast	Error (%)
Month 1/2017	38.280	38.124	-0.407
Month 2/2017	38.590	40.538	5.047
Month 3/2017	39.550	41.004	3.117
Month 4/2017	39.580	39.399	-0.357
Month 5/2017	41.290	42.440	2.784

Forecast for next 3 months

	Month 5/2017	Month 6/2017(f)	Month 7/2017(f)	Month 8/2017(f)	Source
PC_4	1.635	1.460	1.177	1.101	Original variable(f)
PC_5	0.499	0.614	1.090	0.721	Original variable(f)
FPT	**41.29**	**41.451**	**41.147**	**41.068**	[1]
Average error		+/- 1.582	+/- 1.565	+/- 1.561	
% of Average error		+/- 3.82	+/- 3.80	+/- 3.80	[2]
$P_r (Z=1/ PC_1,...,PC_5)$	**0.889**	**0.888**	**0.908**	**0.875**	(f): Dự báo

References

- [1] Modelling of a stock's price forecast in the context of high dimensional data set (Fair 2017)
- [2] Simulation and analysis of forecast risks on high dimensional data set (Fair 2017)

QUESTIONS AND ANSWER ?