ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Mutual Information and Gamma Test for Input Selection

Nima Reyhani, Jin Hao, Yongnan Ji, Amaury Lendasse *

Helsinki University of Technology - Neural Networks Research Center
P.O. Box 5400, FI-02015 - Finland

Abstract. In this paper, input selection is performed using two different approaches. The first approach is based on the Gamma test. This test estimates the mean square error (MSE) that can be achieved without overfitting. The best set of inputs is the one that minimises the result of the Gamma test. The second method estimates the Mutual Information between a set of inputs and the output. The best set of inputs is the one that maximises the Mutual Information. Both methods are applied for the selection of the inputs for function approximation and time series prediction problems.

1 Introduction

Input selection is one of the most important issues in machine learning especially when the number of observations is relatively small comparing to the number of features. Mathematically speaking, a finite set of inputs is sufficient in order to extract an accurate model out of the infinite observations [1]. In practice, there is no data set with infinite number of data points and furthermore, the necessary size of the data set increases dramatically with the number of observations (curse of dimensionality). To circumvent this, one should select the best features or inputs in the sense that they contain the necessary information. Then it would be possible to capture and reconstruct the underlying regularity or relationship between input-output data pairs. With respect to this, some approaches have been proposed, such as branch and bound and Bayesian selection [2-6].

Some of them deal with the feature selection problem as a generalization error estimation problem. In this methodology, the set of features that minimize the generalization error are selected using Leave-one-out, Bootstrap or other resampling technique. These approaches are very time consuming and may take several weeks. However, there are other approaches [7-12] which select *a priori* features based only on the dataset and so the computational cost would be less than the cost of the model dependent cases. Model independent approaches select a set of features by optimizing a criterion over different combinations of inputs. The criteria computes the dependences between each combination of input features and the corresponding output using predictability, correlation, mutual information or other statistics.

Various alternatives for input selection exist. Then, some comparative studies might be helpful as a reference for practical experiments. In this paper, we focus on two promising criteria: a new method called Gamma Test, and, a more conventional

* Part this work is supported by the project of New Information Processing Principles, 44886, of the Academy of Finland.

method, the Mutual Information criterion. The paper is organized as follows: In section 2 and 3 Gamma Test and Mutual Information are introduced. In section 4, LS-SVM is defined in order to compare the inputs selected by each method. In section 5 we present two experimental results (a toy example and a real dataset). Finally, in section 6, conclusions are given.

2 Gamma Test

The Gamma Test (GT) is a technique for estimating the variance of the noise, or the mean square error (MSE), that can be achieved without overfitting [12]. GT is useful for evaluating the nonlinear correlation between two random variables, namely, input and output pairs. It is a generalization of the approach proposed in [12], which is basically based on the fact that the conditional expectation (1) approaches variance of the noise when the distance between the data points tends to zero [13].

$$\varepsilon\left\langle \frac{1}{2}(y'-y)^2 \,\middle|\, |x'-x| < \delta \right\rangle \quad as \quad \delta \to 0 \,. \tag{1}$$

The GT has been applied to various problems in Control Theory, feature selection and secure communication [12]. The experiments show that GT is efficient and thus can be applied to real world problems. A mathematical proof of GT can be found in [13] and it is based on a generalization of Chybechov inequality and the property of k-nearest neighbor structures. In [13], three conditions are necessary:
- the first and second partial derivatives of the underlying function exist;
- the first to the fourth moments of the noise distribution exist;
- the noise is independent to the corresponding points.

Using these three conditions, the variance of the noise is given by the bias term, called Γ, of the regression between $\gamma(k)$ and $\delta(k)$, where $1 \le k \le p$.

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} \left| y_{N[i,k]} - y_i \right|^2 \tag{2}$$

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} \left| x_{N[i,k]} - x_i \right|^2 , \tag{3}$$

With $x_{N[i,k]}$ the k^{th} nearest neighbour of x_i and $y_{N[i,k]}$ the corresponding output. According to [12], $p = 10$ is used in experiments presented in section 5.

3 Mutual Information

The mutual information (MI) of two variables, let say X and Y, is the amount of information obtained from X in the presence of Y, and vice versa. MI can be used for evaluating the dependencies between random variables, and has been applied to Feature Selection and Blind Source Separation [14].
Let's consider two random variables; the MI between them would be,

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

where H(.) computes the Shannon's entropy. Equation (4) leads to integrations and some approaches have been proposed to evaluate them numerically [15]. In this paper,

a recent estimator based on *k*-nearest neighbours statistics is used [16]. The novelty of this approach consists in its ability to estimate the MI between two variables of any dimensional spaces. The basic idea is to estimate H(.) from the average distance to the *k*-nearest neighbours (over all x_i). MI is derived from equation (4) and is estimated as following:

$$I(X,Y) = \psi(k) - \frac{1}{k} - \left\langle \psi(n_x) + \psi(n_y) \right\rangle + \psi(N) \tag{5}$$

where N is the size of data set, k is the number of nearest neighbours and $\psi(x)$ is the digamma function

$$\psi(k) = \Gamma(x) - 1 - d\Gamma(x)/dx \tag{6}$$

and $\psi(1) \approx -0.5772156$.

$$< \ldots >= N^{-1} \sum_{i=1}^{N} E\left[\ldots(i)\right]. \tag{7}$$

$n_x(i)$, $n_y(i)$ are the number of points in the region $||x_i - x_j|| \leq \varepsilon_x(i)/2$ and $||y_i - y_j|| \leq \varepsilon_y(i)/2$, $\varepsilon(i)/2$ is the distance from z_i to its *k*-nearest neighbors, and $\varepsilon_x(i)/2$, $\varepsilon_y(i)/2$ the projections of $\varepsilon(i)/2$ [16]. $k = 6$ is used in the experiments.

4 Least Squares Support Vector Machines

LS-SVM are regularized supervised approximators. Comparing to SVM, it does not have local minima and the optimisation process is simpler. A short summary of the LS-SVM model is given here; more details are given in [17].

The LS-SVM model [17, 18] is defined in its primal weight space by

$$\hat{y} = \omega^T \varphi(\mathbf{x}) + b \tag{8}$$

where $\varphi(x)$ is a function which maps the input space into a higher dimensional feature space, $\mathbf{x}$ is the N-dimensional vector of inputs and x_i, and ω and b are the parameters of the model. In Least Squares Support Vector Machines for function estimation, the following optimization problem is formulated:

$$\min_{\omega,b,e} J(\omega,e) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{9}$$

subject to the equality constraints

$$y^i = \omega^T \varphi(\mathbf{x}^i) + b + e^i, i = 1,\ldots,N. \tag{10}$$

In equation (10), the superscript I refers to the number of the sample. The parameter set θ consists of vector ω and scalar b. Solving this optimization problem in dual space leads to finding the α_i and b coefficients in the following solution:

$$h(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x},\mathbf{x}_i) + b. \tag{11}$$

Function $K(x,x_i)$ is the kernel defined as the dot product between the $\varphi(x)^T$ and $\varphi(x)$ mappings. The meta-parameters of the LS-SVM model are the width of the Gaussian kernels (taken to be identical for all kernels) and the γ regularization factor.

LS-SVM can be viewed as a form of parametric ridge regression in the primal space. The training method for the estimation of ω and b parameters can be found in [17].

5 Experimental Results

The two methods presented in the previous sections are used to select the best input variables (from a set of possible variables) by evaluating the MI or Γ value. All the combinations of input features, e.g. 2^d-1, are tested (d is the number of input variables). Then, the one that gives the maximal MI and the one that give the minimal Γ, are selected. Two experiments on a small datasets have been performed in order to show the level of efficiency of GT and MI for the problem of input selection. The first one is a toy example for which the correct inputs are known. The second used example is a benchmark in the field of time series prediction: the Poland Electricity Dataset [19].

5.1 Toy Examples

In this experiment, we investigate the robustness of Mutual Information and Gamma Test for selecting the correct inputs against additional noises. First we use the following equation for generating a toy data set:

$$Y = X_1 X_2 + \sin X_7 + X_{10} + a * \varepsilon \tag{12}$$

with ε a uniform noise in [-1, 1], X is a uniform distributed 10-dimensional variable, with and a the weighting coefficient of the noise. The number of X observations is 1000 (the size of the dataset). The robustness of each approach is tested by increasing the value of parameter a in order to detect when it fails in the selection of the correct inputs. The results are illustrated in Fig. 1 for both approaches.

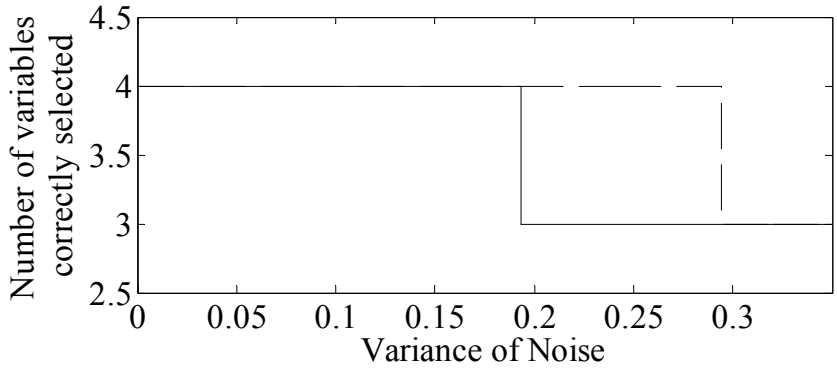

Fig. 1: Toy Example: number of correct selection for GT (solid line), number of correct selection for MI (dotted line).

In this example, GT is less robust than MI in selecting the right inputs in presence of large amount of noise. MI starts to fail when the variance of the noise becomes higher than 0.2940 and GT starts to fail when the variance of noise becomes higher than 0.1933.

5.2 **Poland Electricity Data Set**

In order to predict the next value of a Time Series, an auto-regressive model is used:

$$y(t) = f(y(t-1), y(t-2), ..., y(t-n)) . \quad (13)$$

The input variables in the right-hand part of (13) form the regressor. In our experiment, n is equal to 8. The methodology presented in section 5 is used to select the best regressor: 2^d-1 regressors are tested. The selected regressor based on GT is $\{y(t\text{-}1), y(t\text{-}2), y(t\text{-}5), y(t\text{-}7), y(t\text{-}8)\}$ and the one based on MI is $\{y(t\text{-}1), y(t\text{-}2), y(t\text{-}6)\}$.

Least Square Support Vector Machine (LS-SVM) is used for comparing the regressor selection performances. For each experiment, two thirds of the whole data set has been used for training, and the remaining data points for testing. Leave-one-out procedure for model selection purposes has been applied. The parameters γ and σ for the GT based regressors are 2164.4 and 0.584; and for the MI based regressors are: 10 and 0.1 correspondingly. The mean absolute error (MAE) on the test set is 0.02464 in case of MI based regressor and 0.01944 in case of GT. The corresponding mean square error (MSE) is 0.00163 in MI case and 0.00103 in GT case. For this experiment, the suggested regressor from the GT is more accurate than the MI one. It indicates that the predictions for GT based regressor are closer to the optimal input set.

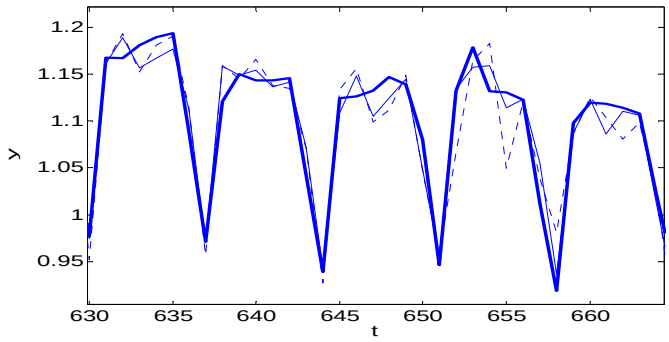

Fig. 2: Test set: true value (thick solid line), prediction based on GT (thin solid line and prediction based on MI (dotted line).

6 Conclusions

In this paper, we presented the efficiency of Gamma Test for feature selection problem on a real dataset and on a toy example. In the literature, it has been demonstrated that MI is a good approach for input selection problems [20] and it has been used as a reference method in this paper.

Based on the experiments, the prediction results obtained with GT are more accurate than the one obtained by MI. But MI is more robust than GT in presence of noises with large variances. The prediction results on the test set show that the input selection based on GT leads to more accurate results than ones based on MI.

References

[1] R. de Figueiredo, Implications and applications of Kolmogorov's superposition theorem, IEEE Tran. on Automatic Control, Vol. 25, Issue 6, p.p. 1227- 1231, Dec 1980.

[2] N. Kwak, CH. Choi; Input feature selection for classification problems, Neural Networks, IEEE Transactions on Volume: 13 , Issue: 1, pp. 143 – 159, Jan 2002.

[3] D. Zongker, A. Jain, Algorithms for feature selection: An evaluation Pattern Recognition, Proceedings of the 13th International Conference on Volume: 2 , 25-29 pp. 18 - 22 vol.2, Aug 1996.

[4] X Chen. An improved branch and bound algorithm for feature selection. Pattern Recognition Letters. Volume 24, Issue 12 (August 2003) Pages: 1925-1933. ISSN: 0167-8655. 2003.

[5] E. P. Xing, M. I. Jordan, R. M. Karp. Feature Selection for High-Dimensional Genomic Microarray Data. Proc. of the Eighteenth International Conference in Machine Learning, ICML2001.

[6] V. Fabio, J. W. Christian. Variational Bayesian Feature Selection for Gaussian Mixture Models. Feature Analysis for ASR, TTS, and Verification. SP-P6.3.

[7] B. Efron, R. J. Tibshirani, An introduction to the bootstrap, Chapman & Hall, 1993.

[8] B. Efron, R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. J. Amer. Statist. Assoc. 92:548–560, 1997.

[9] M. Stone, An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, J. Royal. Statist. Soc., B39, 44–7, 1977.

[10] R. Kohavi. A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Proc. of the 14th Int. Joint Conf. on A.I., Vol. 2, Canada, 1995.

[11] B. Efron, Estimating the error rate of a prediction rule: improvements on cross-validation. Journal of American Statistical Association, 78(382):316–331, 1983.

[12] A. J. Jones, New Tools in Non-linear Modeling and Prediction. Computational Management Science, Vol. 1, Issue 2, p.p. 109-149, 2004.

[13] D. Evans and A. J. Jones, A proof of the Gamma test, Proc. Roy. Soc. Lond. A, Vol. 458, pp. 1-41, 2002.

[14] H. H. Yang and S. Amari, Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information Neural Comput, vol. 9, pp. 1457-1482, 1997.

[15] A. M. Fraser and H. L. Swinney, Phys. Rev. A 33, 1134 (1986).

[16] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information Phys. Rev. E, in press

[17] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002 (ISBN 981- 238-151-1).

[18] J. A. K. Suykens, J. De brabanter, L. Lukas, J. Vandewalle, Weighted least squares support vector machines: robustness and sparse approximation, Neurocomputing, Special Issue on fundamental and information processing aspects of neurocomputing, 48 (1-4), pp. 85-105, 2002.

[19] M. Cottrell, B. Girard, P. Rousset, Forecasting of curves using a Kohonen classification. Forecasting 17, pp. 429–439, 1998.

[20] N. Kwak, C. Choi, Input Feature Selection by Mutual Information Based on Parzen Window,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24 , Issue 12, pp. 1667 - 1671, Dec 2002.