ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

The permutation test for feature selection by mutual information.

D. François[1], V. Wertz[1] and M. Verleysen[2] *

Université catholique de Louvain - Machine Learning Group
1- CESAME Research Center, Av. G. Lemaitre, 4
2- Microelectronics Laboratory, Pl. du Levant 3
Louvain-la-Neuve, Belgium

Abstract. The estimation of mutual information for feature selection is often subject to inaccuracies due to noise, small sample size, bad choice of parameter for the estimator, etc. The choice of a threshold above which a feature will be considered useful is thus difficult to make. Therefore, the use of the permutation test to assess the reliability of the estimation is proposed. The permutation test allows performing a non-parametric hypothesis test to select the relevant features and to build a Feature Relevance Diagram that visually synthesizes the result of the test.

1 Introduction

Selecting features before building a neural model is theoretically useless : the model will set to zero the weights corresponding to unrelevant inputs. However, in practive, it is important to reduce as much as possible the dimensionality of the input space to avoid convergence problems, overfitting, etc.

One possible way to select the features that are relevant for a classification or function approximation problem, is to assign each feature individually a statistical relevance measure, independently from the model subsequently used, and then to select those features that are above a certain threshold. This is often referred to as the *filter* approach, or the *feature ranking* approach.

Mutual information is a non-parametric measure of relevance ; it is derived from information theory. It is powerful (since it is non-parametric) though difficult to estimate (because it is non-parametric). Hence, the estimation of the mutual information can be noisy, unreliable, biased, in cases of small sample size, bad choice of the parameter of the estimator, etc. As a consequence, the choice of a sound threshold is very difficult to make.

To overcome this limitations, the permutation can be applied to the mutual information; this allows (1) selecting a sound threshold resulting from a hypothesis test and (2) detecting bogus estimations of the mutual information.

Section 2 will introduce problem of feature selection, the notion of mutual information and the permutation test. Section 3 will present the combined

*The work of D. François is funded by the Belgian F.R.I.A. M. Verleysen is Research Director of the FNRS. Part of this work is supported by the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

approach of mutual information and permutation test to feature selection ; examples are given in Section 4.

2 Background

2.1 The problem of feature selection

The problem of feature selection is : given $\mathbf{X} = (X_1, \cdots, X_d)$ an input random vector and Y an output random variable, find the subset of indices of the X_i that are most relevant to predict the value of Y [1].

Instead of considering all 2^d possible subsets, we will consider ranking features individually, and choose the k most relevant variables. This approach is sub-optimal with respect to the objective, however it is computationally much less demanding. Its main drawbacks is that (1) it might choose more variables than necessary because it does not take redundancy into account, and (2) it will miss variables that are relevant together although useless individually. Another approach, more elaborate yet still sub-optimal, is to use the mutual information with a forward-backward subset search strategy [2, 3]. This approach is not considered here, although the proposed procedure could be straightforwardly extended to be applied in such situations.

The crucial elements of the feature ranking approach are (1) to estimate the relevance of a feature and (2) to choose the number of features to keep.

2.2 The Mutual Information

The mutual information of two random variables X_i and Y is a measure of how Y depends on X_i and *vice versa*. It can be defined from the entropy $H(.)$:

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) \tag{1}$$

where $H(Y|X)$ is the *conditional* entropy of Y given X_i. In that sense, it measures the loss of entropy (i.e. loss of uncertainty) of Y when X_i is known. If x_i and Y are independent, $H(X,Y) = H(X) + H(Y)$, and $H(Y|X) = H(Y)$. In consequence, the mutual information of two independent variables is zero.

For a continuous random variable X_i, the entropy is defined as

$$H(X) = -\int p_{X_i}(\xi)\, log\, p_{X_i}(\xi)\, \mathrm{d}\xi$$

where p_{X_i} is the probability distribution of X_i. Consequently, the mutual information can be rewritten, for continuous X_i and Y, as

$$MI(X;Y) = \iint p_{X_i,Y}(\xi,\zeta)\, log\, \frac{p_{X_i,Y}(\xi,\zeta)}{p_{X_i}(\xi)\cdot p_Y(\zeta)}\, \mathrm{d}\xi \mathrm{d}\zeta, \tag{2}$$

it corresponds to the Kullback-Leibler distance between $p_{X_i,Y}(\xi,\zeta)$ the joint probability density of X and Y and the product of their respective marginal distributions. In the discrete case, the integral is replaced by a finite sum.

The mutual information can be estimated from two vectors of sample x_i and y from both (1) and (2) using nonparametric density estimation techniques like histograms, kernels, splines or nearest neighbors [4]. It should be noted that all the above mentioned techniques depend on some parameter, like the number of bins for the histogram, the width of the kernel for kernel based density estimation, the number of nearest neighbors for the methods based on nearest neighbors, etc. The choice of that parameter has often to be made 'blindly', that is without any reliability measure for the choice of the parameter. Nevertheless, the estimation can be very sensitive to that parameter, especially in small noisy samples conditions. If some inadequate value is chosen, it can sometimes lead to estimations of the mutual information that are consistently negative for each feature!

2.3 The Permutation test

Although model selection-like approaches (hold out estimates, cross-validation) could be used to test the parameters, we propose to use the permutation test that allows a formal hypthesis test and provides an automatic threshold. The permutation test is a nonparametric hypothesis test [5] over some estimated statistic $\hat{\theta}$ involving x_i and y, which can be a difference of means in a classification context, or correlation, etc. Let $\hat{\theta}_i$ be the value of the statistic for the given x_i and y, both vectors of size n. The aim of the test is to answer the following question : how likely is the value $\hat{\theta}_i$ given the vectors x_i and y if we suppose that they are independent and thus that the statistic θ_i should be zero?

The permutation test considers the empirical distribution of x_i and y to be fixed, as well as the sample size. The random variable of interest is the value of the statistic θ. In such a framework, the distribution of $\hat{\theta}$ is the set of all values of $\hat{\theta}_k$ for all $n!$ possible permutations of the elements of the vector x_i, or, equivalently, all permutations of the elements of the vector y. The P-value α associated to the test is the proportion of $\hat{\theta}_k$ that are larger than $\hat{\theta}_i$.

In practice, the number of all permutations $n!$ can be too large to be tractable; then a subsample of the distribution of θ can be considered ; some permutations are randomly drawn. This is sometimes called the Monte-Carlo permutation test, or the randomized permutation test. In this case, the exact P-value cannot be known ; rather a 95% confidence interval around the observed P-value can be estimated as [6]

$$95\%\text{confidence interval} = \alpha \pm 1.96 \cdot \sqrt{\frac{\alpha(1-\alpha)}{M}}$$

where M is the number of random permutations that are considered.

3 When the permutation meets the mutual information

Fusioning both methods will allow (1) to automatically test the significance of the mutual information and (2) to automatically assess the accuracy of its estimation. In other words, it will help deciding for which variables the measure of mutual information is significantly larger than zero, and how good the estimation of mutual information is.

3.1 The procedure

Let x_i be the vector whose jth component is the value of the ith feature of the jth observation. The length of x_i is n the sample size. The procedure is as follows :

```
1. Choose a significance level $\alpha$.
2. Choose a number of random permutations $M$.
3. Choose a mutual information estimator $mi(\cdot, \cdot; k)$.
4. For each variable $x_i$
5.      Compute $\hat{\theta}_i = mi(x_i, y; k)$
6.      Build $\Theta_i = \{mi(\pi_{x_i}, y; k) | \pi_{x_i}$ is a random permutation of $x_i\}$
7.      Find $\theta_c$ the $100(1-\alpha)$th percentile of the sample $\Theta$
8.      If $\theta_i < \theta_c$ discard the feature
9.      Estimate $\mu_i$ and $\sigma_i^2$ respectively the mean and variance in $\Theta$
```

If the estimated bias (the average of the μ_i) and/or variances (averages of σ_i^2) of the estimator are considered too large compared with the empirical distribution of the mutual information, another estimator is chosen (i.e. another technique or the same technique with another parameter value k), and the procedure is resumed from step 4. The number of permutations M allows to estimate a confidence interval around the significance level.

3.2 The feature relevance diagram

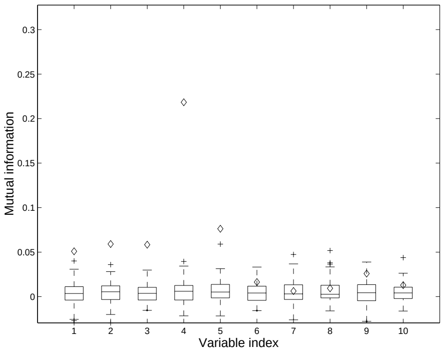
The feature diagram presents in a visual manner the information brought by the permutation test ; it consists in a plot where the horizontal axis represents the variable index, and the vertical axis is the mutual information. For each variable, three elements can be depicted :

1. The value of the mutual information between this variable and the output,
2. A box-plot of the mutual information (from the permutation test)
3. The value of the 100(1-α)th percentile

The value of the mutual information is depicted on Figure 1. as a ‘diamond’. The horizontal edges represent the 25th and 75th percentiles. The bar in the box is the median. The plusses are ‘outliers’ of the distribution.

Another quantity of interest is the value of the mutual information of y with itself. The latter value actually corresponds to the entropy of the variable Y and gives an upper bound on the possible values of the mutual information. If some variable has a mutual information close to that value, then it can be used alone and gives good results for approximation.

The diagram gives a visual representation of the relevance of each variable. All variables for which the mutual information falls into the box-plot of the permutation test can be assumed useless. The relevance diagram furthermore allows to visually estimate the bias and variance of the estimator, respectively with the position of the medians and the sizes of the boxes.



Variable	Mutual information	Critical value	Decision
X_1	0.0509	0.0401	Keep
X_2	0.0591	0.0360	Keep
X_3	0.0583	0.0298	Keep
X_4	0.2184	0.0395	Keep
X_5	0.0762	0.0590	Keep
X_6	0.0163	0.0332	Discard
X_7	0.0063	0.0473	Discard
X_8	0.0093	0.0517	Discard
X_9	0.0259	0.0389	Discard
X_{10}	0.0129	0.0438	Discard

Fig. 1: (and Table 1) Mutual information measurements (diamonds) and thresholds (uppermost 'plus') for the ten variables. Only the first five have been used in the model. As expected, according to the permutation test, they are the only useful ones.

4 Examples

We will consider a synthetic prediction problem, derived from Friedman's [7]. We consider 10 input variables X_i and one output variable Y such that

$$Y = 10 \sin (X_1 \cdot X_2) + 20 (X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon$$

All $X_i, 1 \leq i \leq d$ are uniformly distributed over $[0, 1]$, ϵ is a normal random variable with variance $\sigma^2 = 1$. Variables X_6 to X_{10} are just noise and have no predictive power. Sample size n is 500. The estimation of mutual information is achieved with histogram-based techniques.

4.1 Selecting features automatically

The measured values for the mutual information are given in the second column of Table 1. Except for X_4, all variables seem to have a low mutual information with the output Y. Nevertheless, we know that the first five are used to build the output. The permutation test allows to set a quite precise decision threshold. It was chosen here to be the largest observed value of the permutation test, with $M = 1000$ permutations performed. The P-value of the test belongs to the interval $[0, 0.062]$ with 95% probability.

Table 1 shows that if the decision to keep or discard a given feature according to the test whether the mutual information is larger than the critical value, all the variables included in the model are selected while the others are dismissed.

4.2 Detecting bogus mutual information estimation

We illustrate the use of the permutation test to detect bogus estimation of mutual information, or at least decide for instance which of two estimators to use. To this end, we reduce the sample size to $n = 200$ in order to be able to perceive a difference between estimations made with different histogram bin numbers.

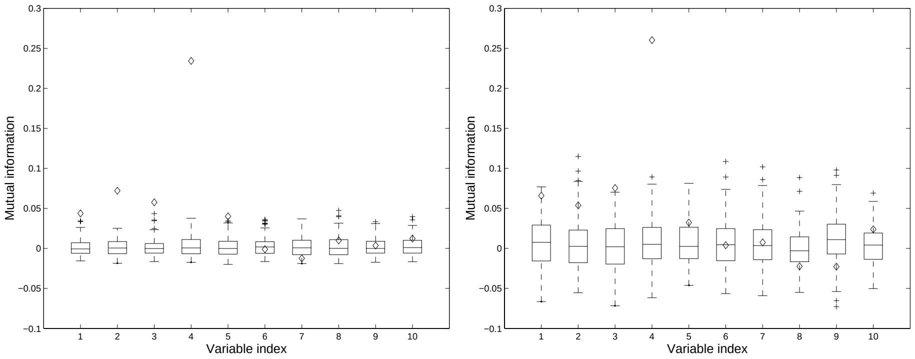


Fig. 2: Relevance diagram for estimation of mutual information with 4 (left) and 10 (right) bins. The variance of the estimator is lower with 4 bins than with 10 bins.

In the previous example, using 10 bins was perfectly fine ; however, with less than half the same number of samples, 10 bins is not optimal. Figure 2 shows the relevance graph for 4 and 10 bins respectively. The variance of the estimator with 10 bins (7.82e-4) is much larger than with 4 bins (1.37e-4), indicating that the 4 bins are more appropriate than 10 bins.

5 Conclusion

The permutation test can be used in conjunction with mutual information to select relevant features for a prediction problem, automatically defining a sound threshold on the value of the mutual information to decide which features to select and which to reject. The permutation test is also usefull to detect when the estimation of the mutual information is not accurate. The procedure could be extended to cope with iterative feature selection procedures based on the mutual information.

References

[1] Isabelle I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[2] R. Battit. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 4(3):537–550, 1991.

[3] A. Kraskov, Harald Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:066138, 2004.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, 1991.

[5] P. Good. *Permutation Tests.* Springer, NewYork, 1994.

[6] J.D. Opdyke. Fast permutation tests that maximize power under conventional monte carlo sampling for pairwise and multiple comparisons. *Journal of Modern Applied Statistical Methods*, 2(1):27–49, May 2003.

[7] J. Friedman. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 9(1):1–141, 1991.