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

Random Forests Feature Selection with K-PLS: Detecting Ischemia from Magnetocardiograms

Long Han[1], Mark J. Embrechts[1], Boleslaw Szymanski[2],
Karsten Sternickel[3] and Alexander Ross[3]

1- Rensselaer Polytechnic Institute - Dept of Decision Sciences &
Engineering Systems, Troy, NY - USA

2- Rensselaer Polytechnic Institute - Dept of Computer Science
Troy, NY - USA

3- Cardiomag Imaging, Inc. Schenectady, NY - USA

Abstract. Random Forests were introduced by Breiman for feature (variable) selection and improved predictions for decision tree models. The resulting model is often superior to AdaBoost and bagging approaches. In this paper the random forests approach is extended for variable selection with other learning models, in this case Partial Least Squares (PLS) and Kernel Partial Least Squares (K-PLS) to estimate the importance of variables. This variable selection method is demonstrated on two benchmark datasets (Boston Housing and South African heart disease data). Finally, this methodology is applied to magnetocardiogram data for the detection of ischemic heart disease.

1 Partial Least Squares (PLS) and K-PLS

Partial Least Squares Regression (PLS) was introduced by Herman Wold [1] for econometrics modeling of multi-variate time series. PLS can be viewed as a "better" Principal Components Analysis (PCA) regression method, where the data are first projected into a different and non-orthogonal basis, and only the most important PLS components (or latent variables) are considered for building a regression model (similar to PCA). The difference between PLS and PCA is that the new set of basis vectors in PLS is not a set of successive orthogonal directions that explain the largest variance in the data, but are actually a set of conjugant gradient vectors to the correlation matrix. The NIPALS implementation of PLS [2] is elegant and fast.

Rosipal introduced K-PLS in 2001 [3] as a nonlinear extension to the linear PLS method instead of using linear kernel K-PLS [4]. This nonlinear extension of PLS makes K-PLS a powerful machine learning tool for classification as well as regression. Powerful variable selection methods have been implemented for PLS and K-PLS, and unlike SVMs, multiple output models are easy to implement. K-PLS can also be formulated as a paradigm closely related (and almost identical) [5] to Support Vector Machines (SVM) [6, 7]. K-PLS uses the same kernel trick as is commonly used in SVMs. K-PLS also provides a purely statistical method, that has been widely used in chemometrics during the past decade. In addition, the idea of using of K-PLS rather than SVMs can be motivated on several levels: (i) PLS is the method by choice in chemometrics and drug design, and K-PLS is a natural extension to PLS; (ii) K-PLS results are generally

comparable to those obtained from SVMs (Table 1); (iii) A powerful feature selection procedure has been implemented with K-PLS that is fully benchmarked and ranked well in the 2003 NIPS feature selection challenge [8].

2 Variable Selection with Random Forests

Dimensionality reduction is a challenging problem for supervised and unsupervised machine learning for classification, regression, and time series prediction. In this paper we focus on variable selection for supervised classification and regression models. The taxonomy of variable selection can be divided into two branches: variable ranking and subset selection [9, 10]. Variable subset selection can be further divided into (i) wrappers, (ii) filters and (iii) embedded methods. The pros and cons of different variable selection methods vary depending on the specific domain problem, computational expense, complexity, and robustness [9]. The motivation of introducing random forests feature selection is to provide an alterative method, which is easily understood, and powerful compared with other feature selection approaches.

Evangelista et al. recently introduced the concept of fuzzy ROC curves and extended this technique to a novel random forests K-PLS modeling technique for variable selection [11]. Random Forests (RF) were introduced by Breiman [12] as a combination of decision tree predictors. RF consist of several hundred models with randomly selected variable subsets (i.e., there is a different subset of training and validation data for each individual model). The main idea is that after generating a vast number of trees, they vote for the most popular variables based on performance. In [12], bagging is used in tandem with RF variable selection in order to reduce the variance. In this paper we extend this random forests idea to estimate the importance of variables with PLS and K-PLS models.

RF variable selection consists of (i) variable subset selection and (ii) aggregate bagging models with variable ranking. For each variable subset a PLS or K-PLS model is used for training and validation. The validation performance is expressed by the q^2 and Q^2 metrics as described in Section 3. For each variable we will add a voting score based on the $(1 - Q^2)^p$ metric for the model in which this variable participated as illustrated in Figure 1. In the formula above, p is a parameter (usually set to 1.3, determined by trial and error).

Lower ranked variables are eliminated based on empirical performance heuristics. This approach can either be done in a greedy way, where variables are selected after applying several bootstraps as illustrated in Fig. 1, or can proceed iteratively, where a few variables are eliminated at a time, and then the entire process is repeated again. Because the procedure as outlined above might lead to discarding significant variables, we introduce a random gauge variable [8, 13], which can either be a uniform or Gaussian (mean 0 and variance 1). A criterion for selecting the relevant variables can now be established by eliminating variables with voting scores below the score for gauge variable.

After the variable selection stage a new K-PLS model is built based on dif-

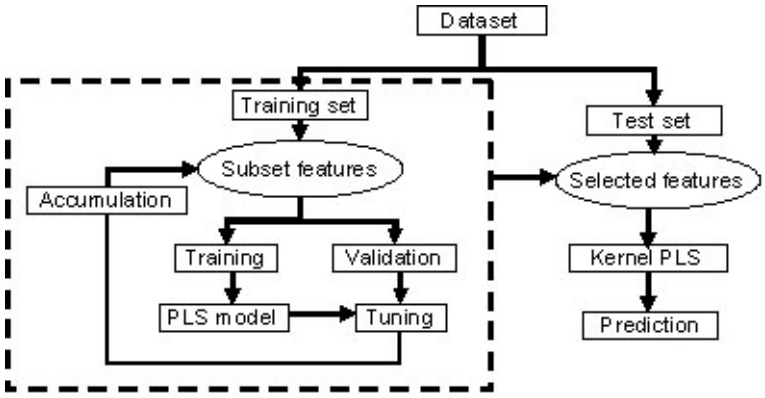

Fig. 1: Model building and validation.

ferent bootstraps with bagging. Predictive models are compared for different variable selection methods based on a Sensitivity Analysis [8] and simple linear kernel PLS models with Z-scores for both Boston housing data and South African Heart disease data.

3 Metrics

Two error measures for the training set can be defined. The correlation coefficient squared between target values and predictions for the response, r^2, is given by:

$$r^2 = \frac{(\sum_{i=1}^{n_{train}}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}))^2}{\sum_{i=1}^{n_{train}}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n_{train}}(y_i - \bar{y})^2}$$

A second and more powerful measure is the so-called "Press r squared" or R^2, because it accounts for the residual error as well.

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{train}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{train}}(y_i - \bar{y})^2}$$

Both metrics are less dependent on the scaling and magnitude of the response value than the Least Mean Square Error (LMSE). For similar purposes, q^2 and Q^2, defined as $1 - r^2$ and $1 - R^2$ respectively, are used to assess the performance of validation or test data. The smaller the q^2 and Q^2 the better; ideally, both values should be close to each other. Detailed information about these metrics is given in [14].

4 Experimental Results

4.1 Benchmark Data

Random Forests variable selection with K-PLS was benchmarked with two data sets: South African Heart Data (SAHD) and the Boston housing market data.

The SAHD are a subset from a larger dataset [15] which defines an almost linear classification problem. It describes a retrospective sample of males in a high-risk heart-disease region of the Western Cape in South Africa. There are roughly two controls per case of CHD. It consists of one response and 9 variables: systolic blood pressure (sbp), cumulative tobacco (tobacco), low density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behavior (typea), obesity, alcohol, and age. A total of 462 samples are included in this data set. The Boston housing data is a standard benchmark regression dataset from the UCI data Repository for Machine Learning [16]. These benchmark data have 506 samples with 12 continuous, one binary variables and one response variable.

Datasets	q^2	Q^2	ROC	LMSE	% Correct	Comments
Boston (K-PLS)	0.129	0.135	(0.967)	3.904	-	LVs =12, $\sigma = 4$
Boston (LS-SVM)	0.122	0.128	(0.963)	3.811	-	$\sigma = 4$
Boston (SVM)	0.133	0.135	(0.971)	3.903	-	$\sigma = 4$
Boston (PLS)	0.260	0.278	(0.934)	5.607	-	-
Heart (K-PLS)	0.760	0.766	0.790	0.426	68.8	LVs = 5, $\sigma = 30$
Heart (LS-SVM)	0.730	0.748	0.812	0.421	68.8	$\sigma = 30$
Heart (SVM)	0.750	0.834	0.794	0.445	71.4	$\sigma = 30$
Heart (PLS)	0.749	0.755	0.797	0.423	67.9	-
MCG (K-PLS)	0.595	0.611	0.855	0.776	82.5	LVs = 5, $\sigma = 4$
MCG (LS-SVM)	0.607	0.622	0.845	0.783	82.5	$\sigma = 4$
MCG (SVM)	0.626	0.651	0.838	0.801	81.7	$\sigma = 4$
MCG (PLS)	0.805	0.957	0.761	0.972	73.3	-
Boston (RF)	0.134	0.142	(0.950)	4.008	-	"zn","age"
Boston (Z-scores)	0.138	0.146	(0.954)	4.071	-	"age","indus"
Boston (SA)	0.127	0.134	(0.965)	3.900	-	"zn","indus"
Heart (RF)	0.762	0.768	0.793	0.426	69.6	"sbp","alcohol"
Heart (Z-scores)	0.762	0.768	0.793	0.426	69.6	"sbp","alcohol"
Heart (SA)	0.785	0.793	0.770	0.433	68.8	"sbp","ldl"
MCG (RF)	0.611	0.621	0.852	0.782	81.7	7 vars deleted
MCG (Z-scores)	0.627	0.637	0.848	0.793	78.3	7 vars deleted
MCG (SA)	0.592	0.604	0.859	0.772	83.3	7 vars deleted

Table 1: Experimental results for three datasets (Upper part with all variables; Lower part with reduction variables)

In each data set, 350 data are randomly chosen as training data with the remaining data are considered test. We use normalization scaling to pre-process the data for both data sets. Random Forests approach is used for variable selection with K-PLS models. After variable selection, the training model is built with a leave-one-out model, and the validation results are based on a bagged model prediction.

In order to validate the experimental results, only training data are used for RF feature selection. In each iteration, we divide the training data into two parts. One part is used for training on the randomly selected variables, the other is used for validation. There are therefore two main parametric choices in the model: the number of random variables and the number of training data. For the Boston housing data, 35, 70, and 105 data are chosen for the validation set

over 3,000 Random Forests models. The number of model variables is set at 4, 6, 8, and 10 respectively. For the South African Heart Data, the same number of validation data are used, but only 1000 Random Forests models are applied due to the smaller number of selected variables. The number of model variables is now set to 4, 6, 7 and 8.

The final selection of ranked variables is relatively insensitive to the selection of the number of variables in the validation data, and to the number of variables used in the individual model selection. Based on the relative variable importance metric for the SAHD data and the variables "alcohol" and "sbp" are dropped for the SAHD data. For the Boston housing data, the proportion of residential land zoned (ZN) and age (AGE) are discarded from the original variables. Note that for both data sets only two features were dropped in order to maintain similar performance metrics for the reduced variable set.

RF variable selection for both benchmark datasets was based on the linear K-PLS model as shown in Table 1. Because leave-one-out validation is used for all training models, the performance metrics have a low variance. Note also that there is no significant difference between the q^2 and Q^2 metrics.

4.2 Binary Classification of Magnetocardiograms (MCG)

The aim of this application is the automated detection of ischemic heart disease for MCG data in order to separate and classify abnormal from normal data sets. The data are from 325 patients consisting of 74 features each.

10,000 Random Forests models are used for 40, 50, 60, and 70 variables respectively. The variable ranking is relatively robust with the number of selected variables in the RF as shown in Table 1. In the final model, the 7 variables with the lowest scores are discarded, maintaining a similar Q^2/q^2 performance as for the original 74 variable model.

In addition, Z-scores variable ranking and Sensitivity Analysis are used as well for each data set. The same number of variables were eliminated in the three variable reduction techniques. For the Boston Housing, South African Heart disease and MCG data, 12, 5 and 5 Latent Variables (LVs) were used. Deleted variables are listed in the last column of Table 1. Table 1 shows that Random Forests results outperform Z-scores ranking and RF are close to those obtained from Sensitivity Analysis. Especially, when a large number of variables is discarded, RF variable selection seems to be superior.

5 Conclusion and Future Work

Benchmark data sets were used to examine a novel variable selection method based on Random Forests and K-PLS and this technique was subsequently applied to magnetocardiogram data with good performance results. Random gauge variables were used to determine which variables to discard by retaining a similar performance metric. Future research will aim to automate the RF variable selection procedure with more robust and less empirical procedures.

6 Acknowledgement

This material is based upon work supported by the National Science Foundation under Award Number 0349589. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] H. Wold. Estimation of Principal Components and related Models by Iterative Least Squares. In P.R. Krishnaiah, editor, *Multivariate Analysis*, pages 391–420. Academic Press, NY, 1966.

[2] H. Wold. Path with Latent Variables: The NIPALS Approach. In H. M. Balock, editor, *Quantitative Sociology:International Perspectives on Mathematical and Statistical Model Building*, pages 307–357. Academic Press, NY, 1975.

[3] R. Rosipal and L. Trejo. Kernel Partial Least Squares Regression in Reproducing Kernel Hillbert Spaces. *Journal of Machine Learning Research*, 2:97–128, 2001.

[4] F. Lindgren, P. Geladi, and S. Wold. The Kernel Algorithm for PLS. *Journal of Chemometrics*, 7:45–49, 1993.

[5] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[6] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[7] K. Bennett and M. Embrechts. An Optimization Perspective on Kernel Partial Least Squares Regression. In J. Suykens et al., editor, *Advances in Learning Theory: Methods, Models and Applications*, pages 227–249. NATO Science Series, Series III: Computer and System Sciences - Vol. 190, IOS Press, Amsterdam, The Netherlands, 2003.

[8] M. Embrechts, R. Bress, and R. Kewley. Feature Selection via Sensitivity Analysis with Direct Kernel PLS. In I. Guyon and S. Gunn, editors, *Feature Extraction*. Springer Verlag, 2005.

[9] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[10] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 1-2:245–271, 1997.

[11] P. Evangelista, M. Embrechts, P. Bonissone, and B. Szymanski. Fuzzy ROC Curves for Unsupervised Nonparametric Ensemble Techniques. Proceedings International Joint Conference on Neural Networks, IJCNN Montreal, Canada, July 31 - August 4, 2005.

[12] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[13] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

[14] M. Embrechts, B. Szymanski, and K. Sternickel. Introduction to Scientific Data Mining: Direct Kernel Methods and Applications. In S. Ovaska, editor, *Computationally Intelligent Hybrid Systems: The Fusion of Soft and Hard Computing*, pages 317–362. John Wiley, New York, 2004.

[15] J. Rousseauw, J. du Plessis, A. Benade, P. Jordann, J. Kotze, P. Jooste, and J. Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64:430–436, 1983.

[16] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI Repository of Machine Learning databases, 1998.