Proc. European Symposium on Artificial Neural Networks, Burges, Belgium, April, 2006, pp. 221-226

Random Forests Feature Selection with Kernel Partial Least Squares: 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 forest 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 Introduction

1.1 Brief Review of Partial Least Squares

Partial Least Squares Regression (PLS) [1] was conceived by the Swedish statistician Herman Wold for econometrics modeling of multi-variate time series. His son, Svante Wold, applied PLS to chemometrics in the early eighties [2] and currently PLS has become one of the most popular and powerful tools in chemometrics and drug design. PLS can be viewed as a "better" principal components analysis method, where the data are first transformed into a different and non-orthogonal basis, similar to Principal Component Analysis (PCA), and only the most important PLS components (or latent variables) are considered for building a regression model (just as in PCA). In this context, S. Wold suggests in hindsight that projection to latent structures would be a more meaningful description for the PLS acronym. 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. PLS regression is one of the most powerful data mining tools for large data sets with many variables with high collinearity. The NIPALS implementation of PLS [3] is elegant and fast.

1.2 Kernel Partial Least Squares(K-PLS)

K-PLS was first described in [4] and applied to spectral analysis in the late nineties. In this paper only linear kernels were considered as a means to speed up the calculation procedure for simple cases with relatively few training samples and a larger number of spectral frequencies. Rosipal introduced K-PLS in 2001 [5] as a nonlinear extension to the linear PLS method.

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) [6] to Support Vector Machines(SVM) [7, 8]. K-PLS uses the same kernel trick as is commonly used in SVMs. K-PLS also provides a natural nonlinear extension to the PLS method, 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)Unlike SVMs, there is no patent on K-PLS; (ii) 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 [9]; (iii)K-PLS can be considered as a traditional neural network paradigm that can handle multiple output nodes.

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 [10, 11]. In variable ranking, each variable is ranked using a metric based on classification or prediction performance for a given outcome. Variable subset selection can be further divided into (i)wrappers, (ii) filters and (iii) embedded methods. **Wrappers** use learning machines as a black box to score different subsets according to their prediction performance. **Filters** are based on a separate preprocessing step to select variables based on rational criteria, which are independent from prediction performance (i.e., the response variables in the learning tasks are not considered). **Embedded** methods are model dependent approaches, which integrate the subset selection approach with the machine learning paradigm. It is therefore specific to the learning algorithm. The majority of examples for this approach are related to a direct object optimization methodology. The pros and cons of different variable selection methods vary depending on the specific domain problem, computational expense, complexity, and robustness [10].

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 [12]. Random Forests (RF) were introduced by Breiman [13] 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). In [13], Random Forests were used with decision tree models by aggregating many tree predictors to obtain an improved predictor. The main idea is that after generating a vast number of trees, they vote for the most popular variables based on performance. In [13], 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 feature selection is a combination of the above introduced (i) variable subset selection and (ii) bootstrapping and variable ranking. The Random Forest technique is applied to evaluate each randomly selected subset of variables based on prediction performance. 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, based on empirical robustness experiments).

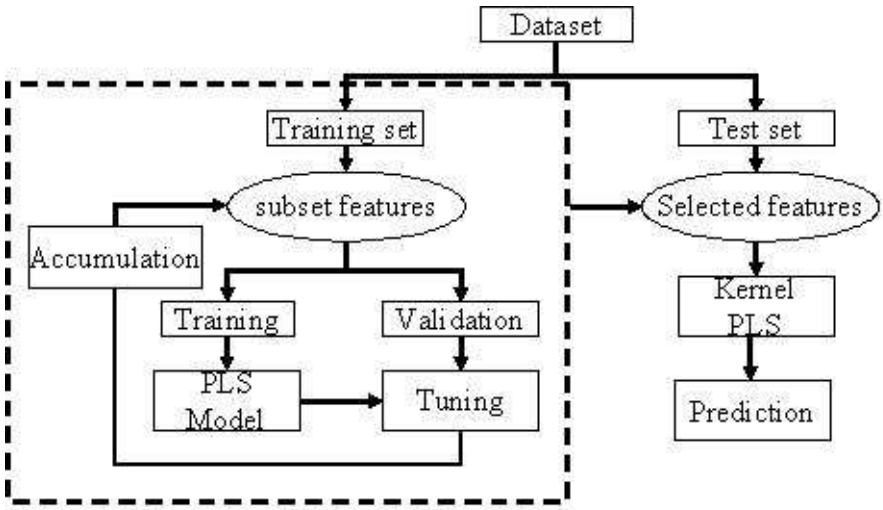

Fig. 1: Model building and validation.

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 uniform or normally distributed random gauge variable [14, 15], which can either be gaussian or uniform (mean

0 and variance 1), and has no relation to the given outcome. A criterion for selecting the relevant variables can now be established by eliminating variables with voting scores below the gauge variable.

After the variable selection stage a new PLS or K-PLS model is built based on different bootstraps with bagging. Predictive models are compared for different variable selection methods based on a sensitivity analysis [15] and simple linear PLS models with Z-scores [16] 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.

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

Both metrics are less dependent on the scaling and magnitude of the response value than the Root Mean Square error. 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 [17].

4 Experimental Results

4.1 Benchmark Data

Random Forest 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 [18] 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 [19]. These benchmark data have 506 samples with 12 continuous, one binary variables and one response variable.

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 Forest approach is used 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 $3,000$ iterations. The number of random 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 iterations are applied due to the smaller number of selected variables. The number of random variables is 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.

RF variable selection for both benchmark datasets was based on the linear K-PLS model as shown in Table 1. 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 Forest 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 SA are used as well for each data set. We eliminate the same number of variables obtained through Z-scores on linear models and compare the two prediction results based on the different variable selection methods. For the Boston Housing, South African Heart disease and MCG data, 12, 5 and 5 latent variables (LVs) are used respectively. Deleted variables are listed on "Comments" column in Table 1. Table 1 shows that Random Forests results outperform Z-scores ranking. Especially, when a large number of variables is discarded, RF variable selection seems to be superior.

Datasets	q^2	Q^2	ROC	RMSE	% Correct	Comments
Boston (Orig)	0.129	0.135	0.967	3.904	-	LVs =12, $\sigma = 4$
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 (Orig)	0.760	0.766	0.790	0.426	68.8	LVs = 5, $\sigma = 30$
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 (Orig)	0.595	0.611	0.855	0.776	82.5	LVs = 5, $\sigma = 4$
MCG (RF)	0.611	0.621	0.852	0.782	81.7	7 vars
MCG (Z-scores)	0.627	0.637	0.848	0.793	78.3	7 vars
MCG (SA)	0.592	0.604	0.859	0.772	83.3	7 vars

Table 1: Experimental results for three datasets

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. Currently, we use random gauge variables to determine which variables to discard. 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. 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.

[2] S. Wold, M. Sjölström, and L. Erikson. PLS-regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130, 2001.

[3] 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.

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

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

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

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

[8] K.P. Bennett and M.J. 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.

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

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

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

[12] 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.

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

[14] 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.

[15] M.J. Embrechts, F.A. Arciniegas, M. Ozdemir, C.M. Breneman, and K.P. Bennett. Bagging neural network sensitivity analysis for feature reduction in QSAR problems. IEEE International Joint Conference on Neural Networks, 2001.

[16] M.H. Kutner, C.J. Nachtsheim, and J. Neter. *Applied Linear Regression Models*. McGraw-Hill Education, 2004.

[17] M.J. 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.

[18] 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.

[19] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998.