Machine Learning algorithms: a study on noise sensitivity

Elias Kalapanidas, Nikolaos Avouris[1]* Marian Craciun[2] and Daniel Neagu[3]

[1] University of Patras, Rio Patras,
GR-265 00, GREECE
[2] University of Dunarea de Jos, Galati,
ROMANIA
[3] University of Bradford, Bradford,
UK

Abstract. In this study, results of a variety of ML algorithms are tested against artificially polluted datasets with noise. Two noise models are tested, each of these studied on a range of noise levels from 0 to 50algorithm, a linear regression algorithm, a decision tree, a M5 algorithm, a decision table classifier, a voting interval scheme as well as a hyper pipes classifier. The study is based on an environmental field of application employing data from two air quality prediction problems, a toxicity classification problem and four artificially produced datasets. The results contain evaluation of classification criteria for every algorithm and noise level for the noise sensitivity study. The results suggest that the best algorithms per problem in terms of showing the lower RMS error are the decision table and the linear regression, for classification and regression problems respectively.

1 Introduction

Noise is a random error of variance of a measured variable [1]. Real datasets coming from monitoring of environmental problems usually contain noisy data, mainly due to malfunctions, unfortunate calibrations of measurement equipment or network problems during the transport of sensor information to a central measurement collection unit. Types of noise are present to almost any real world problem, but not always known.

Predictive algorithms have used synthetic datasets during their development stage. In order to cope with real world problems where the presence of noise in data is a common fact, the algorithms require the existence of a pre-processing module that would deter the impact of noise in data before they are processed. The way this module would calculate the data may significantly affect the performance of the constructed model.

In this study many machine learning algorithms from the machine learning platform Weka [2] are examined in the presence of increasing levels of artificial noise. The gradient impact of noise renders the initial problem from a deterministic one to a stochastic one. The aim of this study is to find machine learning algorithms that would exhibit a

* Research supported ...

good fit to the noise inflicted datasets as well as a smooth degradation as the noise level is increased. The results would be useful for any similar problem facing noise impurities in its data. The algorithms engaged in our study are summed in table 1.

Three sources of data were exploited in this study; two problems of short-hand prediction of daily maximum pollutant concentration, one problem of classification of the toxicity level of various chemical substances, and four artificial datasets were the basis of the research conducted in this paper.

Algorithm	**Weka Scheme**	**Type***	**Description**
Zero Rule	ZeroR	R/C	A very naive algorithm that classifies all cases to the majority class. Used for reference reasons.
K-nn	IB-k**	R/C	The well known Instance-Based algorithm k-nearest neighbors, implemented in accordance to Aha and Kibler [3]
Linear Regression	LinearRegression	R	The linear regression algorithm
M5	M5Prime	R	An algorithm exploiting decision trees with regression on the cases at each leaf
K Star	KStar	R	An Instance-based learner using an entropic distance measure [4]
MLP	Neural Network	R/C***	An implementation of the classical MLP neural network trained by the feed-forward back propagation algorithm
Decision Table		C	A scheme that produces rules formatted as a table, from selected attributes (following a wrapper-type feature selection prior to the training phase)
Hyper Pipes	HyperPipes	C	For each class a HyperPipe is constructed that contains all points of that class. Test instances are classified according to the class that most contains the instance.
C4.5 decision tree	J48	C	An implementation of the C4.5 decision tree [7]
C4.5 Rules	J48.PART	C	A scheme for building rules from partial decision trees
Voting Feature Interval	VFI	C	A simple scheme that calculates the occurrences of feature intervals per class, and classifies by voting on new cases [6]

Table 1. A summary of the machine learning algorithms evaluated in the noise sensitivity study.

*: R for Regression type, C for Classification type of problems.

** : for this study 9 neighbors were chosen for the k parameter after preliminary study with another sample of data not used in the final experiments.

***: an MLP with fixed parameters was used, having 20 hidden neurons, sigmoid activation function on each neuron, 500 epochs of training with a learning rate of 0.2 and a momentum of 0.2.

2 Past experience on noise sensitivity of machine learning algorithms

At the past a small number of noise studies have been reported, as for the sensitivity of machine learning algorithms against this problem. Noise models have been examined on different variants of the TD reinforcement learning algorithm at 1994 [11], as well as on induction learning programs by Chevaeyre and Zucker [5]. Following the tracks of the pioneering idea of Kearns [9] about statistical query models, Teytaud [12] theoretically explains the relation between some noise models and regression algorithms.

Li et al. [10] presented a study about four machine learning algorithms, i.e. a C4.5 type of decision tree, a naive bayes classifier, a decision rules classifier and the OneR method of one rule, on a noise model. They consequently compared the results of the algorithms before and after the wavelet denoising technique, for small levels of noise, finding that the technique applied boosted the efficiency of the algorithms in almost all of the noise levels.

3 Emulating noise: The Noise Model Examined

In the following study, a noise model is applied on the datasets at hand, introducing a white noise type of deformation on the original data. Two assumptions are considered to be true for all datasets:

1. The variables of the dataset (both the independent and the dependent variables) are normally distributed
2. Noise is randomly distributed and independent from the data.

Then for every case (y_i, x_i) in the dataset L: The pair (y_i, x_i) of the dependent variable Y and the matrix of independent variables X is substituted by another $(y_i^{'}, x_i^{'})$ by a probability of n, where n is the noise level. The new pair is calculated by the following formula:

$$x_{ij}^{'} = \begin{cases} x_{ij} + \sigma_{xj} z_j & p_{ij} >= n, \\ x_{ij} & p_{ij} < n. \end{cases} \quad y_i^{'} = \begin{cases} y_i + \sigma_y z_j & p_{ij} >= n, \\ y_i & p_{ij} < n. \end{cases} \tag{1}$$

$z_{ij} = norminv(p_{ij}), j \in [1, .., k]$

σ_{xj} is the standard deviation of x_j, z_j is a normally distributed random variable and is calculated by the inverse function of density-probability of the normal distribution for a value of p_{ij}, having a mean value of zero and a standard deviation equal to unity, $p_{ij} \in (0, 1)$ is a probability variable produced by a random value generator following a uniform distribution.

4 Experiments and discussion

Our experiments are based on two air quality problems, a toxicity classification problem and four artificial datasets. The dependent variables for the first two problems correspond to the maximum daily concentration of nitrogen dioxide and of ozone, two harmful aerial pollutants, after 10 o' clock in the morning. The datasets under study contain five and eight main input attributes respectively that were selected after a feature-selection procedure using a genetic algorithm [8]. The toxicity problem refers to the estimation of the toxicity index of several chemical substances, containing 20 features and 1 dependent variable. Finally, 4 artificial problems have been included in the study as described in [13], consisted of four features and one output variable. Of the four latter datasets, one implements a multivariate problem, another one a linear function, while the third and the fourth ones refer to a non-linear function. Table 2 summarizes this information per problem studied.

Problem Code	**Description**	**Dependent Variable**
A1	Artificial problem	Numerical/ Multivariate
A2	Artificial problem	Numerical/ Linear
A3	Artificial problem	Numerical/ Non linear of the form x2
A4	Artificial problem	Numerical/ Non linear of the form x2
NO2	Daily Maximum Concentration Forecasting from sensory data	Numerical/ Non linear
O3	Daily Maximum Concentration Forecasting from sensory data	Numerical/ Non linear
TOX	Classification of an index of toxicity for various substances from chemical descriptors	Numerical

Table 2. A description of the problems of the noise-sensitivity study.

The artificial problems A1-A4 are of the type $y = f(x_1, x_2, x_3, x_4)$ and have been created using the formulas of table 3:

Problem	**x1=**	**x2=**	**x3=**	**x4=**	**y=**
A1	z	$x_1 * 0.8 + z * 0.6$	$x_1 * 0.6 + z * 0.8$	z	$(x_1 - x_2 - x_3 + x_4)/1.47$
A2	z	$(x_1^2 + z * 0.5)/1.5$	$x_1 * 0.6 + z * 0.8$	z	$(x_1 - x_2 - x_3 + x_4)/1.7$
A3	z	$x_1 * 0.8 + z * 0.6$	$x_1 * 0.6 + z * 0.8$	z	$(x_1 - x_2^2 - x_3 + x_4)/1.96$
A4	z	$x_1 * 0.8 + z * 0.6$	$x_1 * 0.6 + z * 0.8$	z	$(x_1 - x_2 - x_3 + x_4^2)/1.76$

Table 3. Definition of the variables for the four artificial problems A1-A4.

All cases containing missing values have been deleted. Although the resulting datasets may already contain an amount of noise, for this study it should be considered as clean data and this fact does not influence the experiments as the noise models studied refer to additive noise.

Artificial noise was generated at random throughout the whole datasets. The reason of "polluting" both the training set and the evaluation set is that as noise in the data has been emerged so far, it will be emerged in the future with the same probability and the same patterns.

Repetitive experiments have been done on each of the polluted datasets. Eight noise levels have been tested, ranging from 0 to 500.20, 0.30, 0.40, 0.50}. Five different datasets were produced for every such noise level, while the results of the competing machine learning schemes were averaged over these five datasets. Five-fold cross validation experiments were carried out for each of these datasets. For each run, eleven machine learning algorithms were trained and tested following the five-cross-validation scheme. These are summed in table 1 along with a short description for each one of them. Half of them are suitable for regression type of problems, while the other half is classification algorithms. Since all the problems were regression problems, the numerical dependent variable for each of the problems was transformed into a categorical one by dividing the initial range into 5 equally wide areas. Thus the seven initial regression datasets were transformed into seven classification datasets, ready to be processed by the classification type of algorithms.

From the variety of the collected data from this study the metric of RMS error was chosen to judge the fit of each machine learning algorithm to every artificially polluted dataset. Though other metrics as the prediction accuracy or the classification error are also used in many publications, the RMS error is a stricter and more suitable evaluator of the efficiency of a certain algorithm in terms of a comparison study.

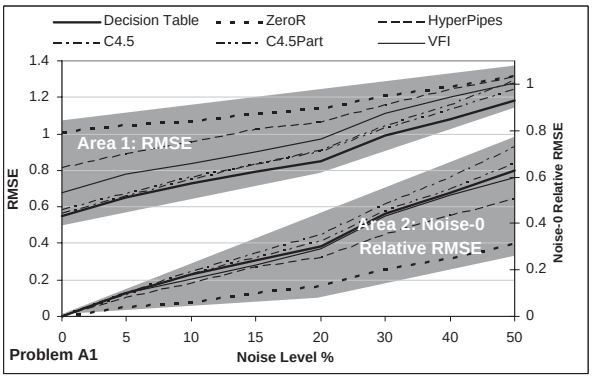

Fig. 1. Example of a noise sensitivity diagram: RMSE and Noise-0 Relative RMSE areas.

In the next diagrams the sensitivity of the algorithms to noise is depicted, by the form of noise curves. Each of the 7 problems is represented by a pair of diagrams,

one for the regression type and one for the classification type of algorithms. For every diagram the left vertical axis represents the RMS error of the algorithms while on the right vertical axis the noise-0 relative RMS error is measured. The latter error metric is the difference of RMS error after the application of noise in data minus the RMS error at the zero noise level. On the horizontal axis the gradually increasing levels of artificial noise reside. The less sensitive algorithms to the presence of noise are those that follow a noise-0 relative RMSE line as close to the horizontal axis as possible.

For readability and space compactness purposes it was chosen to have both RMSE and noise-0 relative RMSE curves in the same diagram. These two curves are utilizing different areas, as the example of fig.1 indicates. Area 1 containing the RMSE curves always takes over the upper - upper left part of the diagram, while area 2 where the noise-0 relative RMSE fits in is contained in the lower - lower right part of the diagram.

4.1 Inquiring the regression results

It is clear from the diagrams referring to the noise curves of the regression type algorithms that their behavior varies from the artificial problems to the real-world problems.

In all of the four artificial problems A1-A4 the best algorithms show a very good fit to the noise free problems at 0 noise level, but after that level their RMS error jumps up abruptly and then evolves almost linearly. This behavior is more visible by watching the noise-0 relative RMSE curves. For the first two linear problems A1 and A2 linear regression proves to be the best method as expected, followed closely by M5. For the two non-linear problems A3 and A4 M5 and IB-9 fit better. Though these algorithms appear to have the better RMSE curves, their corresponding noise-0 relative RMSE curves are among the worst. It emerges as a conclusion from the four problems A1-A4 that the weaker algorithms appear as the less sensitive to noise, and vice versa.

The real-world problems NO2, O3 and TOX present a different image. All RMSE curves are gathered inside a narrow band, close to each other. In all cases linear regression fits better the datasets of these problems. In disagreement with the conclusion from the artificial problems A1-A4, the noise-0 relative RMSE curves mirror the same behavior of their RMSE counterparts.

4.2 Exploring the classification results

For all the problems except TOX, VFI and HyperPipes are the less fit algorithms, having RMSE curves over the reference algorithm ZeroR. Since ZeroR is used as an efficiency threshold, all algorithms exhibiting RMSE curves over its own are considered unsuitable for solving the problem at hand. From the other three algorithms, Decision Table is the dominant one, having the lowest RMSE curve and the lowest noise-0 relative RMSE for all the seven problems studied in this report.

Another interesting finding is that the classification algorithms have noise curves much less sensitive than those of the algorithms of the regression type. This may be a result of the discretization of the dependent (output) variable. We assume that as the number of the discrete bins of the discretization process increases, the average slope of the noise curves will also increase so as to match this of the regression type of algorithms when the number of bins reaches the infinity.

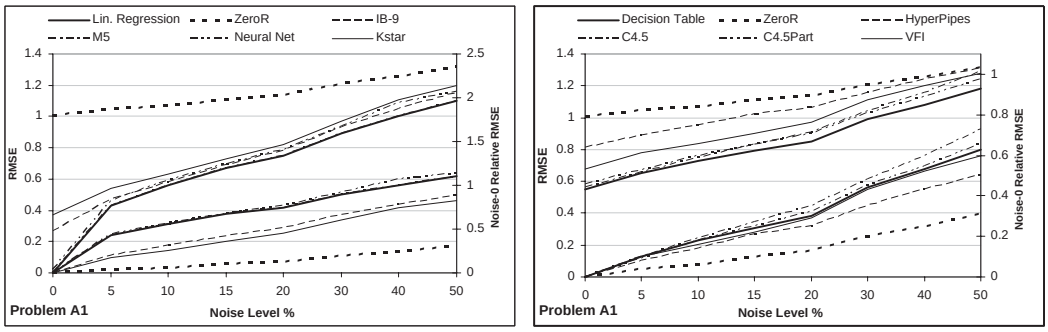

Fig. 2. RMS and noise-0 relative RMS Error for the A1 Problem.

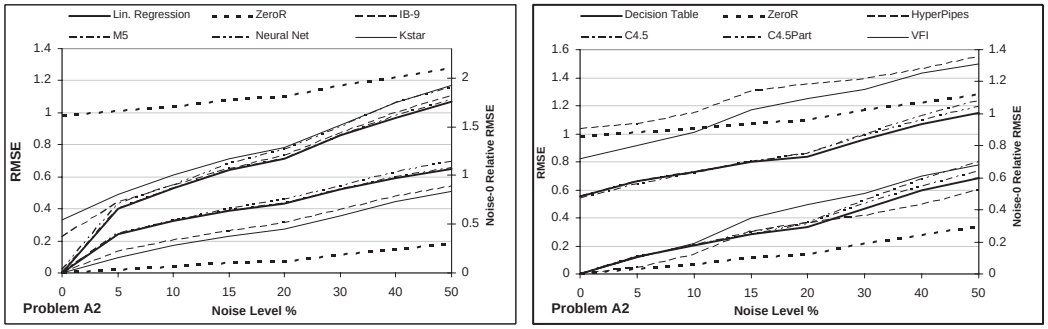

Fig. 3. RMS and noise-0 relative RMS Error for the A2 Problem.

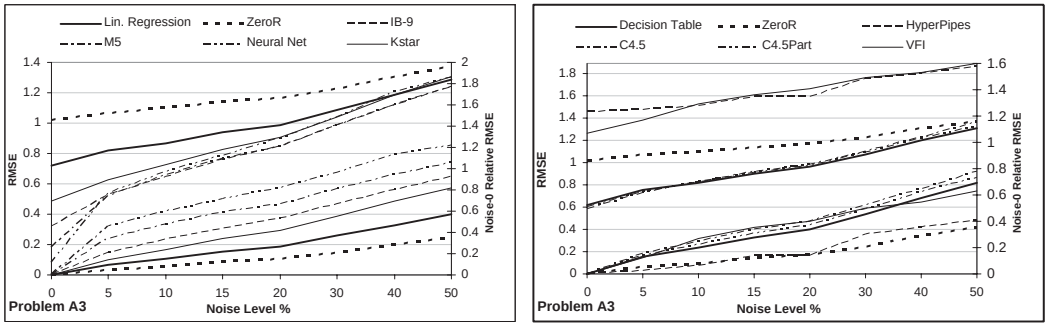

Fig. 4. RMS and noise-0 relative RMS Error for the A3 Problem.

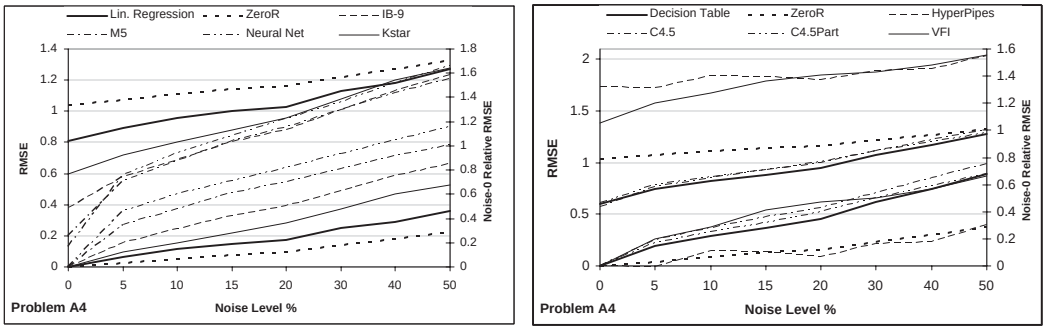

Fig. 5. RMS and noise-0 relative RMS Error for the A4 Problem.

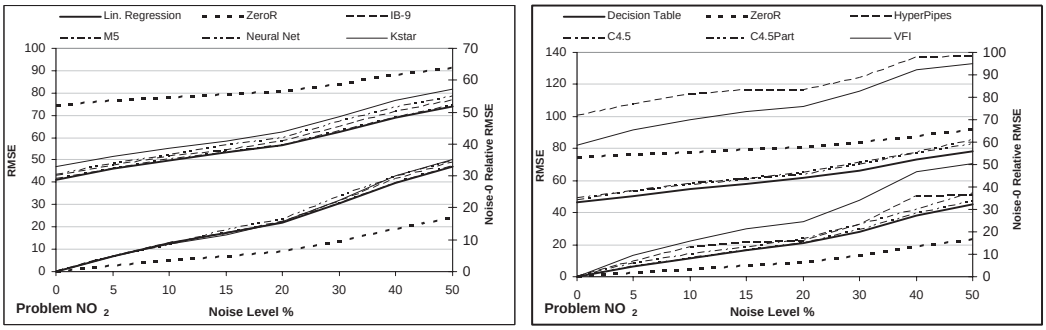

Fig. 6. RMS and noise-0 relative RMS Error for the NO2 Problem.

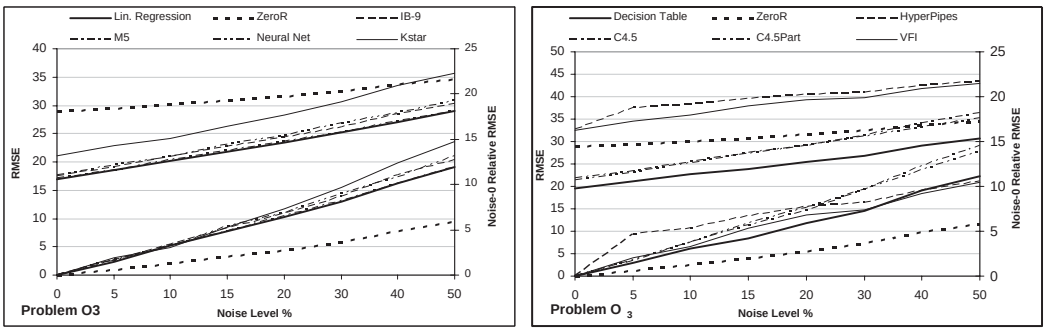

Fig. 7. RMS and noise-0 relative RMS Error for the O3 Problem.

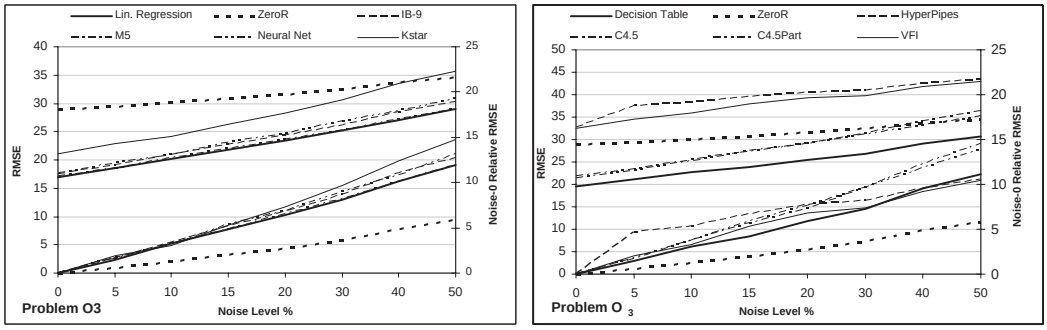

Fig. 8. RMS and noise-0 relative RMS Error for the TOX Problem.

5 Conclusions

A study of machine learning algorithms on noise sensitivity is reported in this paper, based on four artificial and three real world problems. A noise model has been tested, for a noise level ranging from 0 to 0.5. The dependent variable was transformed from numerical to nominal in order to test the classification algorithms. Thus a range of regression type and classification type of algorithms have been examined by measuring their sensitivity to noise, as artificially additive noise has been applied on the initial datasets.

It has been showed that linear regression from the regression type of algorithms adapts better to the gradually increasing noise levels. Also noticeable from the artificial datasets A1-A4 is the fact that the better algorithms in terms of RMSE present the greater noise sensitivity while the worst seem to be the less sensitive. Decision Table seems to be the method the less sensitive to additive noise from the set of classification learners. Not only does it show the best RMSE on all of the datasets, but exhibits a good behavior in terms of the noise-0 relative RMSE.

Future work expanding the reported study includes further experiments on different problems, and we believe that the forthcoming results will help in forming general guidelines useful for the selection of the best machine learning algorithm for modeling or prediction of problems prone to noise. At last it is worth noting that all data for the examined datasets are originated from PERPA, the Greek air quality monitoring authority.

References

1. Han J. And Kamber M., "Data Mining: Concepts and Techniques". Morgan Kaufmann Publishers, 2000
2. Garner, S.R., (1995), WEKA: The Waikato Environment for Knowledge Analysis. In Proc. of the New Zealand Computer Science Research Students Conference, pages 57-64.
3. Aha, D., and D. Kibler (1991), "Instance-based learning algorithms", Machine Learning, vol.6, pp. 37-66

4. John, G. Cleary and Leonard, E. Trigg (1995) "K*: An Instance- based Learner Using an Entropic Distance Measure", Proceedings of the 12th International Conference on Machine learning, pp. 108-114.
5. Chevaleyre Y., Zucker J-D., (2000), "Noise-Tolerant Rule Induction for Multi-Instance Data". In Proceedings of the ICML-2000 Workshop on "Attribute-Value and Relational Learning".
6. Demiroz, G. and Guvenir, A. (1997) "Classification by voting feature intervals", ECML-97
7. Quinlan J.R., (1993), "C4.5, Programs for Machine Learning", Morgan Kauffman, San Mateo, California, USA.
8. Kalapanidas E. and Avouris N., (2002), "Feature Selection Using a Genetic Algorithm Applied on an Air Quality Forecasting Problem". In Proceedings of the BESAI (Binding Environmental Sciences with AI) Workshop, ECAI 2002, Lyon, France.
9. Kearns M., (1993), "Efficient noise-tolerant learning from statistical queries". In Proceedings of the twenty-fifth annual ACM symposium on Theory of computing, p.392-401, May 16-18, San Diego, California, United States .
10. Li Q., Li T., Zhu S., Kambhamettu C., (2002), "Improving Medical/Biological Data Classification Performance by Wavelet Preprocessing", In the Proceedings of ICDM 2002.
11. Pendrith M., (1994), "On reinforcement learning of control actions in noisy and non-Markovian domains". Technical Report UNSW-CSE-TR-9410, School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia
12. Teytaud O., (2001), "Robust Learning: Regression Noise". In Proceedings of IJCNN 2001". pp 1787-1792.
13. Sarle, W.S. (1998), "Prediction with Missing Inputs," in Wang, P.P. (ed.), JCIS 98 Proceedings, Vol II, Research Triangle Park, NC, 399-402, ftp://ftp.sas.com/pub/neural/JCIS98.ps.