Random Brains: An ensemble method for feature selection with neural networks

Mark J. Embrechts[1], Jorge M. Santos[2] and Jonathan D. Linton[3] *

1- Rensselaer Polytechnic Institute - Dept. of Industrial and Systems Engineering
Troy, NY - USA

2- School of Engineering, Polytechnic of Porto - Dept. of Mathematics, and
Biomedical Engineering Institute, Porto - Portugal

3- University of Ottawa - Telfer School of Management
Ottawa - Canada

Abstract. The purpose of this paper is to introduce and validate Random Brains, a novel artificial neural network based feature selection technique. Feature selection is widely used in high-dimensional data and it aims on removing irrelevant or redundant data, providing faster predictors without a significant decrease in model performance. Random Brains, inspired by Breiman's Random Forests, are bagged ensembles of predictive neural network models that use randomly selected subsets of features. This paper validates Random Brains on several classification and regression benchmark data sets by comparing its performance to similar models with features selected based on sensitivity analysis.

1 Introduction to feature selection

Feature or variable selection is an important step in understanding and explaining the performance of a predictive model. The terms feature selection and variable selection can be use interchangeably in this paper, although some authors reserve the terminology features for newly made up (latent) variables from the original (manifest) variables [7].

Because of interdependencies and inter-correlations between variables, there is no unique set of reduced variables that best explains a predictive model. Variable selection generally aims to identify a reduced set of variables that allows for the formulation of predictive models with little loss or in some cases better performance accuracy. While such a reduced set of variables is not unique, preference is usually given to a compact set of variables that also make sense to the domain expert and that allow for the formulation of easy to understand explanative rules for the model.

2 Random Brains

Random Brains refers to an artificial neural networks ensemble method for variable selection inspired by Breiman's random forests [2]. Random forests are a combination of decision tree predictors that uses randomly selected features or

*The authors acknowledge the support of the National Sciences and Engineering Research Council (NSERC) of Canada in conducting this research.

combinations of features as inputs. There are several advantages to methods based on the random forests concept such as: (i) efficiency on large data sets; (ii) generation of an unbiased estimation of the generalization error; (iii) estimation of the relative variable importance; (iv) robustness to outliers and noise; and (v) simple and straightforward parallelization of the algorithm.

In the practical implementation of the Random Brains algorithm the number of neural networks in the ensemble (K) and the number of randomly selected features for each ensemble subset (R) need to be pre-specified. The neural networks are trained by early stopping (by using a validation set consisting of 10 % of the training patterns). A key issue with random brain models is to let neural networks train to completion without human oversight. This is achieved by setting layer-specific learning parameters following the procedures outlined in LeCun's Efficient BackProp [11] and using additional refinements reported in [3]. The validation performance metric for each neural network in the Random Brains ensemble is captured via the Q^2 metric (explained in Section 4): a tally of the relative feature importance is updated by adding $1 - Q^2$ on the tally for the selected features. This procedure leads to a hierarchy of features based on their relevance or relative performance as depicted in Figure 1. Useful information can be obtained from this chart: if the chart is *flat* features are equivalently relevant; if the chart is not *flat* there are features with different levels of relevance and the chart can guide us on selecting the number of features for the final model.

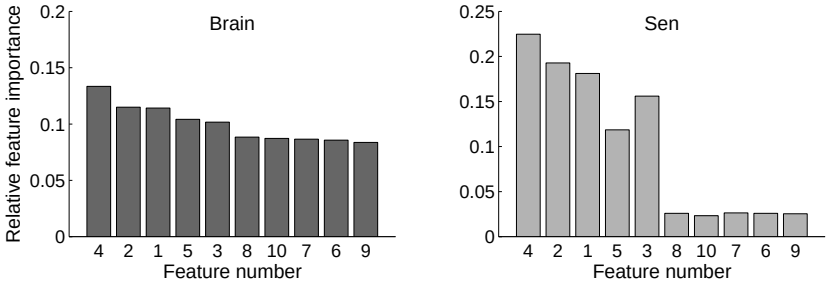

Fig. 1: Feature relevance with Random Brains and sensitivity analysis

The final stage consists of choosing the top L features to use in the final neural network model. The selection of L can be based on different strategies depending on the problem and on the shape of the feature relevance chart. The feature selection procedure can proceed either iteratively or greedy. A detailed analysis of various benchmark problems did not show any performance advantage of iterative feature selection over greedy feature selection. Therefore all variable selection procedures in this appear are based on a greedy approach. Because the purpose of this paper is to validate the performance of the Random Brains method, we did not attempt to identify an optimal subset of features. A comparison is made on the predictive performance of models with different subsets of features (90 percent of the features, and 100, 30 and the actual number of relevant features) to similar models using a feature selection methodology

based on sensitivity analysis [10].

3 Data sets

Six benchmark data sets were used to assess the Random Brains feature selection performance (3 binary classification problems and 3 regression problems). Theses data sets were selected because they cover a wide range with respect to the number of data and/or the number of features.

Friedman data set [5] is a synthetic regression data set composed of 5 relevant features (uniformly distributed random variables $N(0,1)$ and 5 additional Gaussian distributed variables represented by ϵ). The response y is related to the 5 uniformly random variables according to $y = 10\sin(\pi x_1 x_2) + 20(x_3 + 0.5)^2 + 10x_4 + 5x_5 + \epsilon$. A total of 1000 data were generated (800 training data and 200 test data).

Boiling Point data [4] are QSAR (Quantitative Structure Activity Relationship) regression data used for the prediction of the boiling point of 298 molecules, 30 of which are used for test data. Typically for such QSAR data sets about 30 variables can be isolated to make good predictive models.

Abalone data [12] consists of 4177 patterns, originally with 8 attributes to predict the age of abalones. This particular data set was augmented with 500 uniformly distributed random numbers in $[0,1]$.

Leukemia data set [6] consists of microarray expression data for 7129 genes to distinguish between two different kinds of leukemia. The first 38 samples are used for training and the last 34 samples for testing. A good binary classifications can often be made based on a selection of 30–100 genes.

Arcene data [8] are based on mass spectrometry data to distinguish cancer patients from normal patients and were augmented with additional random features for the NIPS 2003 feature selection competition. There are 10000 features and the data set is divided into 100 training data and 100 test data.

Advertising data set [1] is used as a binary classification problem to remove internet advertisements from web-based images. There are 3279 data with 1558 features in 2 unbalanced classes (2821 non-ads and 458 ads). The data were randomly split and 2500 data were used for training.

4 Performance metrics

We use a number of performance metrics to evaluate the predictive models, as no one metric sufficiently quantifies model quality. For regression, we start with the mean absolute error (MAE) and the root mean squared error (RMSE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where y_i and $\hat{y}_i$ are the true and predicted responses of the i^{th} data. Note that because these quantities are data dependent, they have no intrinsic value

unless the regressor is Mahalanobis-scaled or standardized. Pearson's correlation coefficient $r^2(y, \hat{y})$ is used to assess the strength of the linear relationship between observed and fitted response values. R^2 is used to assess how well predicted values lie on the main diagonal and is computed as:

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

where $\bar{y}$ is the average response. For reasonable prediction models, r^2 and R^2 are often very similar. We restrict the use of r^2 and R^2 to training data predictions, and $q^2 = 1 - r^2_{test}$ and $Q^2 = 1 - R^2_{test}$ are used to assess test data predictions.

We also use a series of metrics for assessing classification results. The balanced percent correct (BPC) quantifies the average classification rates between the positive and negative classes. The area under the ROC curve (AUC) [14] is used to assess binary classifications. This metric can be extended to multi-class classification and regression tasks, and we use this metric when comparing regression benchmarks as well. Hubert and Arabie's Adjusted Rand Index (ARI) ranges over $[0, 1]$ and assesses the randomness of the classification; the reader is directed to [9] for a detailed explanation of this metric.

5 Comparative study and discussion

Table 1 shows the performance metrics for the three regression data sets for neural network models with all features and reduced numbers of features obtained from Random Brains and sensitivity analysis. The Random Brains model consists of 200 models with a random set of selected features that consist of 50% of the original features. All neural network models have two hidden layers (with 23 and 11 neurons respectively) and were trained using early stopping, where the stopping point is determined using a validation set (10% of the training data). The number of features listed in Table 2 consists of the original number of features, 90% of the original number of features and depending on the data set 100 features, 30 features and the final relevant number of features. For the Friedman data both sensitivity analysis and Random Brains resulted in the same 5 final features (Fig. 1), and the relative strengths of features 1 and 2 obtained with Random Brains are consistent with their symmetric appearance in the equation that generated the data. For the boiling point and abalone data the features selected with sensitivity analysis lead to neural networks with a superior predictive performance, but the performance metrics with features selected with Random Brains are close.

Table 2 shows the performance metrics for the three classification data sets for neural network models with all features and reduced numbers of features obtained with Random Brains and sensitivity analysis. The followed procedures for feature selection and neural network modeling were identical to those used for the regression data. It should be noted that for the arcene data the Random

Brains feature selection method outperformed sensitivity analysis, while the reverse was true for the advertising data and the leukemia data. Contrary to the regression data, there is a significant difference in performance metrics for the data sets for which sensitivity analysis gives a better classification metric.

Data Set	Method	# feats	q^2	Q^2	AUC	$RMSE$	MAE
Friedman	All	10	0.092	0.093	0.961	1.453	1.137
	RB	**5**	**0.054**	**0.055**	**0.983**	**1.115**	**0.872**
	Sen	**5**	**0.054**	**0.055**	**0.983**	**1.115**	**0.872**
Boiling Point	All	184	0.022	0.027	1.000	11.372	8.358
	RB	166	0.026	0.032	1.000	12.574	9.433
	RB	100	0.016	0.017	0.996	9.196	6.720
	RB	30	0.016	0.018	0.996	9.474	7.241
	RB	10	0.023	0.024	1.000	10.889	8.026
	Sen	166	0.030	0.033	0.991	12.768	9.581
	Sen	100	0.023	0.026	0.996	11.193	7.645
	Sen	**30**	**0.013**	**0.015**	**1.000**	**8.684**	**6.767**
	Sen	10	0.036	0.036	1.000	13.287	9.753
Abalone	All	508	0.750	0.782	0.767	2.932	2.115
	RB	100	0.632	0.649	0.812	2.670	1.947
	RB	30	0.582	0.583	0.829	2.530	1.825
	RB	8	0.494	0.495	0.852	2.333	1.649
	Sen	100	0.558	0.563	0.823	2.486	1.791
	Sen	**30**	**0.574**	**0.576**	**0.841**	**2.517**	**1.774**
	Sen	8	0.469	0.471	0.861	2.276	1.565
	Sen	8	0.469	0.471	0.861	2.276	1.566

Table 1: Performance metrics for feature selection with Random brains and Sensitivity analysis for the regression data sets.

Data set	Method	#feats	Q^2	AUC	BRC	ARI	$RMSE$	MAE
Arcene	All	10000	0.53	0.89	81.74	0.31	0.73	0.48
	RB	**9000**	**0.45**	**0.91**	**85.39**	**0.44**	**0.67**	**0.39**
	RB	100	0.86	0.79	71.59	0.24	0.92	0.53
	Sen	9000	0.57	0.91	84.01	0.68	0.75	0.40
	Sen	100	0.84	0.74	69.16	0.10	1.03	0.65
Advertising	All	1558	0.15	0.98	96.26	0.85	0.26	0.07
	RB	1402	0.18	0.99	94.19	0.82	0.29	0.09
	RB	100	0.48	0.90	80.69	0.65	0.47	0.21
	Sen	1402	0.16	0.98	95.41	0.86	0.26	0.06
	Sen	**100**	**0.20**	**0.98**	**94.97**	**0.84**	**0.30**	**0.06**
Leukemia	All	7129	0.46	0.96	90.00	0.52	0.67	0.26
	RB	6111	0.60	0.94	81.43	0.40	0.76	0.36
	RB	100	0.47	0.93	82.14	0.38	0.67	0.40
	RB	30	0.85	0.85	71.43	0.23	0.91	0.53
	RB	10	0.84	0.64	68.93	0.09	1.01	0.65
	Sen	6111	0.64	0.94	83.93	0.40	0.79	0.34
	Sen	**100**	**0.22**	**0.99**	**92.86**	**0.58**	**0.46**	**0.26**
	Sen	30	0.35	0.98	86.79	0.60	0.58	0.28
	Sen	10	0.77	0.89	76.07	0.31	0.86	0.45

Table 2: Performance metrics for feature selection with Random Brains and sensitivity analysis for the classification data sets.

6 Conclusions

This paper introduces and validates Random Brains, a novel neural network based ensemble feature selection method. For some data sets Random Brains led to better selected features when compared to sensitivity analysis. Future research could be focused on determining the optimal number of models and features for Random Brains, establishing proper scaling factors for the feature tally, and developing guidelines to establish the final number of relevant selected features.

References

[1] C. Blake, E. Keogh and C. Merz, UCI repository of machine learning data sets. Available at: `www.ics.uci.edu/~mlearn/MLRepository.html`, 1998.

[2] L. Breiman, Random forests. *Machine Learning*, Vol. 45, pp. 5–32, 2001.

[3] M. J. Embrechts, B. Hargis, and J. D. Linton, An Augmented Efficient Backpropagation Training Strategy for Deep Autoassociative Neural Networks. *Proceedings of the 18th European Symposium on Artificial Neural Networks (ESANN 2010)*, pp. 141-146, April 22-24, 2010. Bruges, Belgium, 2010.

[4] L. M. Egolf, M. D. Wessel, and P. C. Jurs, Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. *Journal of Chemical formation and Computer Science*, Vol. 34. pp. 947–956, 1994.

[5] J. Friedman, Multivariate adaptive regression splines. *Annals of Statistics*, Vol. 19. pp. 1–141, 1991.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lande, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286, pp. 531–537.

[7] I. Guyon and A. Eliseeff, An introduction to variable and feature selection. *Journal of Machine Learning research*, Vol. 3, pp. 1157–1182, 2003.

[8] I. Guyon, S. Gunn, A. B. Hur, and G. Dorr, Design and Analysis of the NIPS 2003 Challenge. In I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (eds.), *Feature Extraction: Studies in Fuzziness and Soft Computing*, Vol. 207, pp. 237–263, Springer, 2006.

[9] H. L. Hubert and P. Arabie, Comparing partitions. *Journal of Classification*, Vol. 2, pp. 193–218, 1985.

[10] R. Kewley, M. Embrechts, and C. Breneman, Data strip mining for the virtual design of pharmaceuticals with neural networks. *IEEE Transactions on Neural Networks*, Vol. 11 (3) , pp. 668–679, 2000.

[11] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, Efficient backprop. In G. B. Orr and K. R. Müller, Eds., *Neural Networks: Tricks of the Trade*, pp. 9-50, Springer, 1998.

[12] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford, The population biology of abalone (Haliotis species) in Tasmania. 1. Blacklip abalone (H. Rubra) from the north coast and the islands of Bass Strait. Technical Report, Sea Fisheries Division, Dept. of Primary Industry and Fisheries, Tasmania, 1994.

[13] A. Prinzie and D. Van den Poel, Random multiclass classification: Generalizing random forests to random MNL and random NB. In R. Wagner, N. Revell, and G. Pernul (eds.), *Database and Expert Systems Applications*, Lecture Notes in Computer Science, Vol. 4653, pp. 349–358, Springer, 2007.

[14] J. A. Swets, R. M. Dawes, and J. Monahan, Better decisions through science. *Scientific American*, Vol. 283, pp. 82–87, 2000