International Journal of Software Engineering & Computer Sciences (IJSECS)
ISSN: 2289-8522 ,Volume 1, pp. 75-84, February 2015

DOI: http://dx.doi.org/10.15282/ijsecs.1.2015.6.0006

EVALUATING THE EFFECT OF DATASET SIZE ON PREDICTIVE MODEL USING SUPERVISED LEARNING TECHNIQUE

A. R. Ajiboye, R. Abdullah-Arshah, H. Qin and H. Isah-Kebbe

Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang,
26300 Gambang, Pahang, Malaysia
Email: ajibraheem@live.com

ABSTRACT

Learning models used for prediction purposes are mostly developed without paying much cognizance to the size of datasets that can produce models of high accuracy and better generalization. Although, the general believe is that, large dataset is needed to construct a predictive learning model. To describe a data set as large in size, perhaps, is circumstance dependent, thus, what constitutes a dataset to be considered as being big or small is vague. In this paper, the ability of the predictive model to generalize with respect to a particular size of data when simulated with new untrained input is examined. The study experiments on three different sizes of data using Matlab program to create predictive models with a view to establishing if the size of data has any effect on the accuracy of a model. The simulated output of each model is measured using the Mean Absolute Error (MAE) and comparisons are made. Findings from this study reveals that, the quantity of data partitioned for the purpose of training must be of good representation of the entire sets and sufficient enough to span through the input space. The results of simulating the three network models also shows that, the learning model with the largest size of training sets appears to be the most accurate and consistently delivers a much better and stable results.

Keywords: Prediction, Neural Network, Supervised Learning, Data mining, Data size.

INTRODUCTION

The rationales for developing a prediction model is primarily aimed at predicting for a new record, object, or whatever value assumed by the target attribute based on the input attributes. The relationship between the target attribute and the other attributes are learned from a set of data in which the target attribute is already known, this learning method is usually referred to as supervised learning. Prediction is described in (Skillicorn, 2007) as producing an appropriate label or categorization for new objects, given their attributes, using information gleaned from the relationship between attribute values and labels of a set of example objects.

Developing a predictive model requires using a reasonable size of the dataset for modelling, but what constitutes a reasonable size of data remain vague. In several cases, especially when the data points fall within a few ranges of values such as categorical data, the use of small size of sample data is expected to perform well if the appropriate techniques are used.

There are several techniques used in modelling dataset for the purpose of making predictions, neural network is one of the successful techniques due to its processing capabilities. It is also the technique used in this study to develop the predictive models. The technique has been reported to have performed well in the construction of predictive model in several studies

such as (Özel and Karpat, 2005); (Oladokun et al., 2008); (Bandyopadhyay and Chattopadhyay, 2007) and (Raghuwanshi et al., 2006). Neural network and other widely used techniques are based on fitting a curve through the data, which mainly involve finding a relationship from the predictors to the predicted.

Also, (Rajaraman and Ullman, 2012), describes modelling of the data as simply the answer to a complex query about it. The present study modelled datasets in order to determine the effect of using dataset of various sizes for model construction and some of the questions the paper would answer include: What differences does it make to use small or large dataset to develop a predictive model? How can the error associated with a particular predictive model or the accuracy of a model be determined and what can be done to improve the accuracy or generalization of a predictive model?

In order to address these questions, the study uses a supervised learning technique of neural network to experiment on three different sizes of dataset in order to create network models capable of predicting the target attribute based on the relationships established from the input attributes. Findings from this study unveils the direct effect of model construction using different sizes of the dataset.

This paper is organized as follows: In the next Section, the learning process of Neural Networks is discussed. This is followed by some predictive learning models reported in the literature and the Section that follows presents the proposed approach and the performance of the training process is graphically illustrated. The experimental results are represented and discussed in a separate Section, while the study is concluded in the Section that follows.

NEURAL NETWORKS LEARNING PROCESS

Artificial Neural Networks (ANNs) are a massively parallel and distributed processor that is made up of simple processing units and has a natural propensity for storing experiential knowledge and making it available for use (Haykin, 2009). Back propagation algorithm is one of the most common algorithm used for training in neural network; feed-forward neural network technique requires this algorithm for training. The algorithm adapted from (Han et al., 2012), follows the learning process represented in Figure 1.

ANNs can learn new associations, patterns and functional dependencies; the learning changes the network's memory either by updating its status or by adding new facts and since ANNs do not use a mathematical model of how a system's output depends on its input, they behave as model-free estimators (Suh, 2012).

Input:

D, a dataset consisting of the training tuples and their associated target values;
L, the learning rate;
Network, a multilayer feed forward network.
Output: A trained network
Method:

1. Initialize all network weights and biases ;
2. **while** terminating condition is not satisfied{
3. **for** each training tuple X in D{
4. // propagate the input forward:
5. **for** each input layer unit j {
6. $O_j = I_j$ // output of an input unit is its actual input value
7. **for** each hidden or output layer unit j {
8. $I_j = \sum_i w_{ij}O_i + \theta_j$; //compute the net input of unit j with respect to the previous layer, i
9. $O_j = \frac{1}{1+e^{-I_j}}$; } // compute the output of each unit j
10. // Back propagate the errors:
11. **for** each unit j in the output layer
12. $Err_j = O_j (1-O_j)(T_j-O_j)$; // compute the error
13. **for**each unit j in the hidden layer,
14. $Err_j = O_j (1-O_j)\sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, k
15. **for** each weight w_{ij} in network {
16. $\Delta wij = (l)\, Err_j O_i$; // weight increment
17. $w_{ij} = w_{ij} + \Delta wij$; } // weight update
18. **for** each bias θ_j in network {
19. $\Delta\ \theta_j = (l)\, Err_j$; // bias increment
20. $\theta_j = \theta_j + \Delta\ \theta_j$; } // bias update
21. }}

Figure 1. Back propagation algorithm

RELATED WORKS

There have been several studies on developing predictive models due to its importance, but little or no attention is paid to the effect of using a particular size of the dataset for model construction. The act of developing a predictive model for the purpose of making useful predictions from the dataset is one of the data mining tasks and the concept of data mining relies much on model construction to predict data (predictive mining) or describe data (descriptive mining).

The use of predictive models have helped in decision making by using the information at hand to predict the future. To show that small dataset can be useful for modelling, the study in (McArthur et al., 2013) proposed a method that allows researchers and practitioners to structure

a small amount of data in a way which aids understandings and allows predictions to be made. The study in (Dobbin and Simon, 2007) develops probability models and the paper proposed the sample size determination for prediction in the context of high-dimensional data that captures variability in both steps of predictor development. Predictive models are found useful in exploring the educational data as reported in (Osmanbegović and Suljić, 2012); (Schumacher et al., 2010) and (Bidgoli et al., 2003) for the predictions of student's performance.

Statistical methods for estimating dataset size requirements and classification of microarray data using *learning curves* is proposed in (Mukherjee et al., 2003). The study focuses on the use of the existing classification results to estimate dataset size requirements for future classification experiments. The study also evaluates the gain in accuracy and significance of classifiers built with additional data. The paper reported that, the subsampling procedure gives more accurate estimates of the quantiles of the true error of a classifier as the number of subsample increases.

The study in (Basavanhally et al., 2010) uses an inverse power-law model of statistical learning to predict classifier performance when only a limited amount of annotated training data is available, the paper, however recommends that, results of classifier comparison made on small data cohorts should not be generalized as holding true when large amounts of data become available.

In the study proposed in (van der Ploeg et al., 2014), the modern modelling techniques is perceived as being hungry of data. The study suggested the use of these techniques in medical prediction problems if there is availability of very large data sets with many events. The study opined that, only little is known about the sample size that is needed to generate a prediction model with a modern modelling technique that outperforms the traditional regression based modelling techniques in medical data.

In the present paper, we carried out a number of experiments aimed at developing models using varied sizes of datasets. The resulting outputs of simulating all the trained network models using the same size of an untrained dataset and further computations of the mean absolute error to determine the associated errors in each of the predictive models are analyzed and discussed.

THE PROPOSED APPROACH

In this section, the development of predictive models based on different sizes of the dataset and the evaluation of these models is presented. Three predictive models are developed and evaluated in a Matlab software environment. The students data collected for the purpose of constructing and evaluating the models are transformed to numeric values to make them suitable for the model construction. The dataset has five input attributes and a target output. Generating a model using the technique of neural network involves mapping all the significant patterns and relationships that exist among specified input attributes to predict the target output. How each model is able to achieve this is what the present study focuses on.

This technique uses supervised learning, as the target is provided for the training sets. The study modelled 3 segments of datasets using similar configurations (see Table 1) and similar network architecture as illustrated in Figure 2.

EXPERIMENTATIONS

This study experiments on different sizes of the dataset. Models are constructed based on the size specified in each of the experiments and the target outputs are known (supervised learning).

Experiment I: In the first experiment, 400 datasets are used to construct the predictive model. The technique of feed-forward technique trained the data sets using back propagation algorithm. By default, this technique partitioned training set to 60%, while the leftover data sets are partitioned to the validation set and testing set in equal percentage. While the training set learn the relationship between the input attributes and the target, the validation set track the error that emanates in the course of the training. Training continues until the validation sets triggers the end of training when the error begins to rise. This is to avoid over-fitting. The Mean Square Error (MSE) is the error computed during this process using the formula in (1):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \qquad (1)$$

where n is the number of samples, $\hat{Y}_i$ is the network output and Y_i is the target value.

The transfer function in the output layer controls the network outputs. Both hidden layer and output layer uses the same transfer function as shown in the network architecture (see Figure 2). The testing set, then evaluate the performance of the trained network.

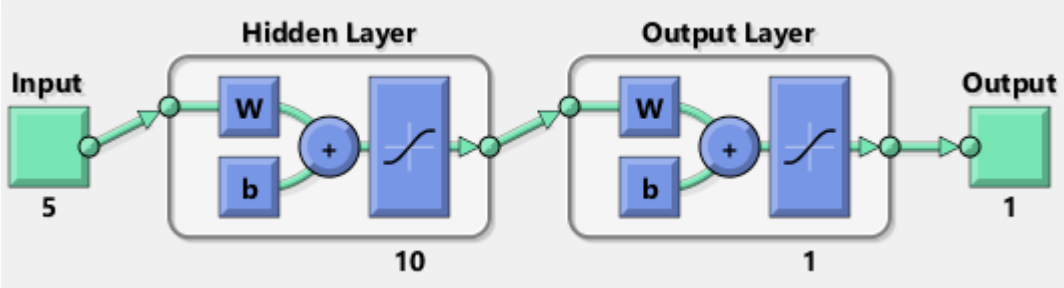

Figure 2. The architecture of the trained network model

Table 1. Network Configurations

Algorithm	*Data Division*: Random *Training:* Levenberg Marquardt
Network properties	*Network type:* Feedforward BP *Performance:* Mean Square Error *Number of Neurons*: 10 *Transfer Function:* TANSIG
Network Parameters	*Epochs*: 750 *Goal*: 0 *Min grad*: 1e-7 *Max_fail*: 6 *Mu* : 0.001

Experiment II: In the second experiment, the datasets used for the model construction are increased to 800. Similarly, using the same technique as in Experiment I, this dataset is partitioned into training set, validation set and testing set. It is worth mentioning here that, while

training and validation sets are needed during the training process, the testing is only needed after the training process converges. The validation sets perform the role as described in Experiment I.

Experiment III: In the third experiment, the data sets used for model construction are increased to 1200. The partitioning of dataset to three portions also conforms to 60%, 20% and 20% for training, validation and testing, respectively. These sets of data, perform roles as described in Experiment I.

TESTING OF THE NETWORK MODEL

The last experiment conducted was to test the network model, in order to determine how the model would respond to a new set of data that the model has not previously seen. This is otherwise referred to as simulation. This testing is quite different from the testing earlier mentioned, which uses an equal percentage as the validation data and only occurred at the convergence of the training process.

Experiment IV: This is the last experiment conducted to simulate each of the predictive models created in the previous experiments using 260 untrained datasets. Simulating each network model gives a network output that is very close to the target output. In order to know which of the models give the closest output to the target, the mean absolute error is computed based on (2):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|F_i - Y_i\right| \qquad (2)$$

where F_i is the prediction from the network model and Y_i is the target value. The results of the errors of the simulated outputs are shown in Table 2. The Mean Absolute Error (MAE) is computed for each predictive model in order to ascertain the pattern of error associated with each model. The choice of MAE among the performance measures for numeric prediction is due to the fact that, MAE does not tends to exaggerate the effect of outliers (Witten et al., 2011), the MAE treats all sizes of error evenly according to their magnitude.

EXPERIMENTAL RESULTS

In this section, the training performance of the three predictive models is illustrated. As shown in the graph, see Figures 3-5, the number of epochs is computed against the Mean Square Error (MSE). Each model converges when the validation sets records an increase in the value of MSE. For instance, if the MSE in the current iteration (E_n) is greater than the error of the previous iteration (E_{n-1}), to avoid over-fitting, the validation set interrupts the training and the network immediately converges. The training performance of the network model with 400, 800 and 1200 datasets is shown in Figures 3, 4 and 5 respectively.

In Figure 3, the error value decreases up to epochs 84 where best performance was recorded and the network converges. Up to this point, the network did not show any sign of over-fitting.

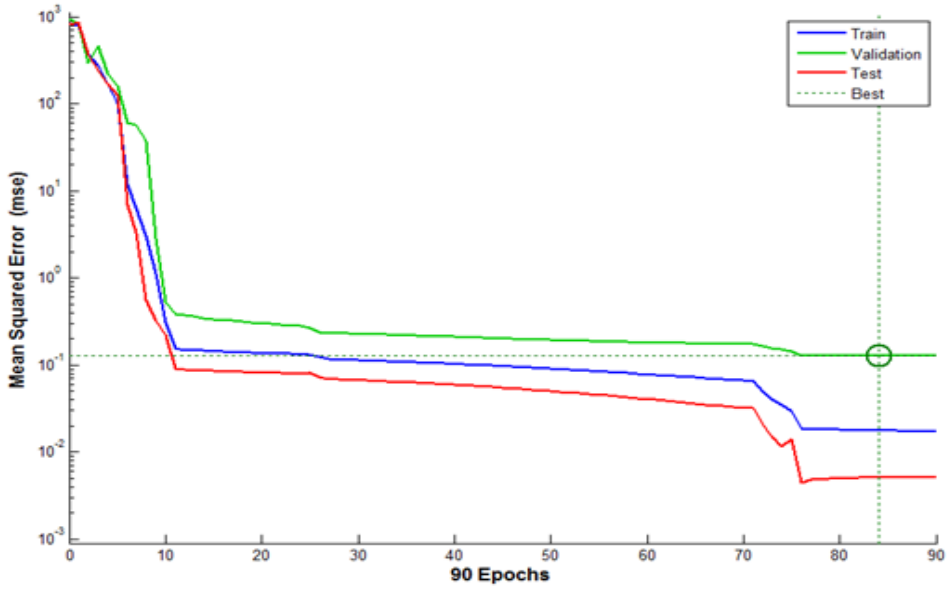

Figure 3. Performance model based on experiment I

In Figure 4, as the number of data sets increase to 800, the training and testing sets depicts more similarity features as shown in the graph; in order word, the error value decreases and no over-fitting is recorded up to epochs 120 where the network converges.

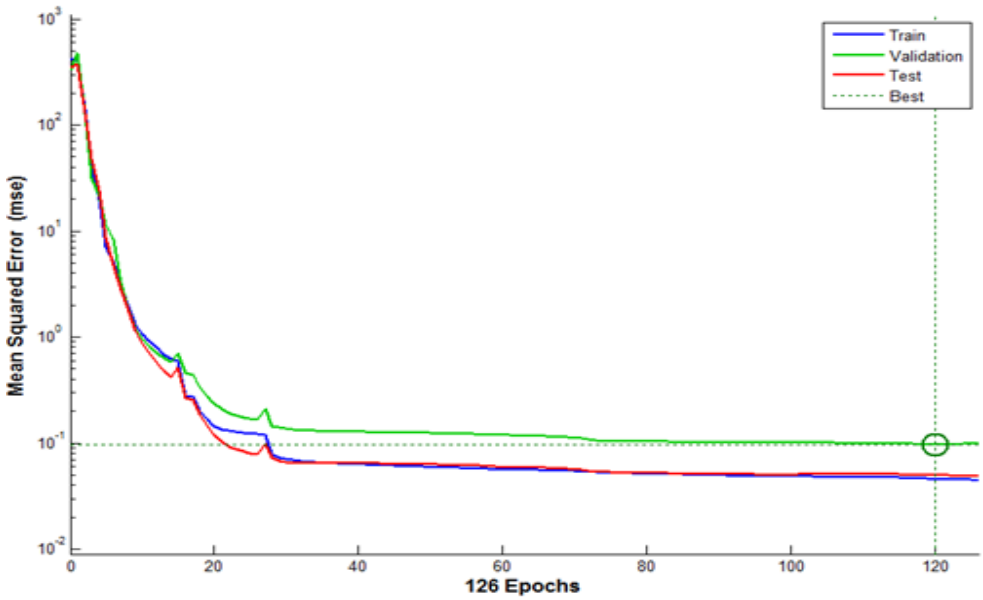

Figure 4. Performance model based on experiment II

The third network model constructed using 1200 data sets is illustrated in Figure 5. The network appears to have been trained so fast here, as the training, validation and testing sets shares many similarities and the network converges and recorded the best performance at epochs 75.

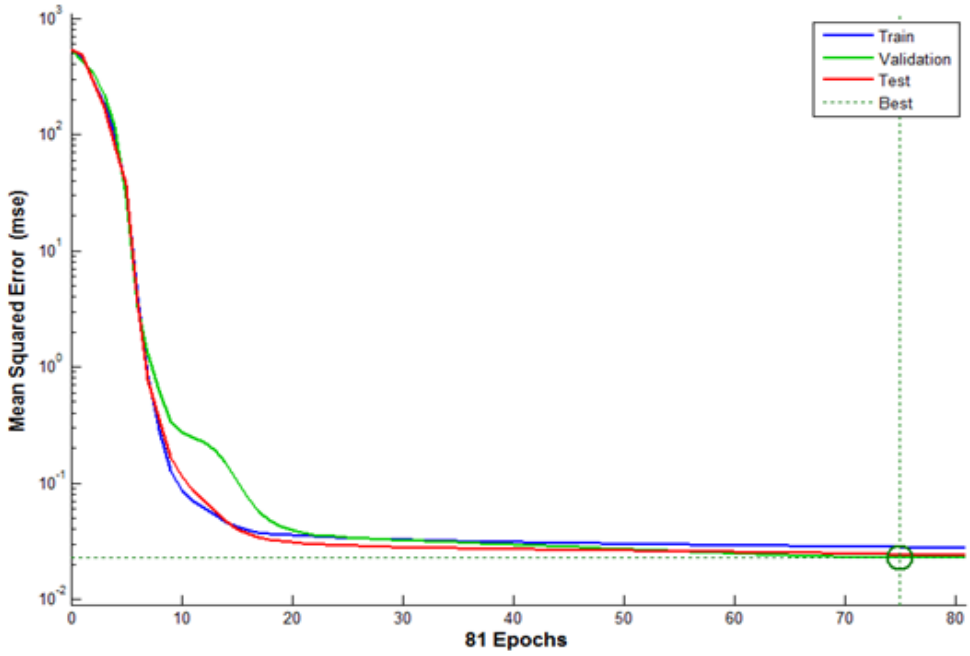

Figure 5. Performance model based on experiment III

The simulated results of each model using a set of untrained inputs are further measured by computing the MAE as shown in Table 2.

Table 2. Mean Absolute Errors of the simulated outputs

Predictive models	The training data sets	Datasets for simulation	Error
Model1	400	260	0.82137
Model2	800	260	0.60431
Model3	1200	260	0.51429

DISCUSSION

In this section, the experimental results are discussed. The performance of the three network models is represented in Figures 3, 4 and 5. From these diagrams, it can be seen that the final mean-square error for Figures 4 and 5 appears very small, while the testing and validation error depicts many similar characteristics. It can also be inferred from the model in Figure 3 that, there are obvious dissimilarities among the training, testing and validation sets. In general, all the models trained well as no over-fitting is shown up to iteration 75. The response of each model to an untrained input dataset, however, gives much better clarifications on their level of accuracies. The results of simulation shown in Table 2 reveal that, model 3 (constructed using 1200 datasets) has the least MAE, while model 1 (constructed using 400 datasets) shows the MAE of highest value. These results have shown that, simulating a predictive model using over 50% of an untrained input data can affect the prediction accuracy. In other word, the training sets should always be large enough to span through the input attributes.

CONCLUSIONS

This paper evaluates and presents the resulting outputs of modelling different sizes of the dataset for prediction purposes.In the course of experimentations, three different sizes of the dataset are modelled using neural network techniques. The study further evaluates the accuracy of each network model by simulating them with a new untrained dataset. The accuracy of the simulated outputs is measured using mean absolute error and the comparison made on these outputs show the degree of error associated with each trained network model.

The predictive model constructed with the smallest dataset records highest error when simulated with untrained inputs, while other models constructed using more dataset records better accuracy. Thus, it can be inferred from the results of this study that, using sufficient data set for predictive model construction can lead to better accuracy and the model's ability to generalize. Although, due to vagueness that surrounds the size of the dataset, it is difficult to say precisely when a dataset can be considered to be big; the results from this study, have shown that, what is the most important is to construct models with adequate size of a dataset that is sufficient enough to span through the input space.

ACKNOWLEDGEMENTS

The authors would like to thank the three anonymous reviewers for their constructive comments. We also thank the Universiti Malaysia Pahang for providing all the facilities required for this study.

REFERENCES

Bandyopadhyay, G., & Chattopadhyay, S. 2007. Single hidden layer artificial neural network models versus multiple linear regression model in forecasting the time series of total ozone. International Journal of Environmental Science & Technology, 4, 141-149.

Basavanhally, A., Doyle, S., & Madabhushi, A. Year. Predicting classifier performance with a small training set: Applications to computer-aided diagnosis and prognosis. Paper presented at the Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, 2010.

Bidgoli, B. M., Kashy, D., Kortemeyer, G., & Punch, W. Year. Predicting student performance: An Application of data mining methods with the educational web-based system Lon-Capa. Paper presented at the Proceedings of ASEE/IEEE frontiers in education conference, 2003.

Dobbin, K. K., & Simon, R. M. 2007. Sample size planning for developing classifiers using high-dimensional DNA microarray data. Biostatistics, 8(1), 101-117.

Han, J., Kamber, M., & Pei, J. 2012. Data Mining Concepts and Techniques (3rd ed.): Morgan Kaufman, Elsevier Inc. USA.

Haykin, S. 2009. Neural Networks and Learning Machines (r. Ed. Ed.). New Jersey: Pearson Education, Inc.

McArthur, D. P., Encheva, S., & Thorsen, I. 2013. Predicting with a small amount of data: An application of fuzzy reasoning to regional disparities. Journal of Economic Studies, 41(1), 2-2.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., . . . Mesirov, J. P. 2003. Estimating dataset size requirements for classifying DNA microarray data. Journal of Computational Biology, 10(2), 119-142.

Oladokun, V., Adebanjo, A., & Charles-Owaba, O. 2008. Predicting students' academic performance using artificial neural network: A case study of an engineering course. The Pacific Journal of Science and Technology, 9(1), 72-79.

Osmanbegović, E., & Suljić, M. 2012. Data mining approach for predicting student performance. Economic Review, 10(1).

Özel, T., & Karpat, Y. 2005. Predictive modeling of surface roughness and tool wear in hard turning using regression and neural networks. International Journal of Machine Tools and Manufacture, 45(4), 467-479.

Raghuwanshi, N., Singh, R., & Reddy, L. 2006. Runoff and sediment yield modeling using artificial neural networks: Upper Siwane River, India. Journal of Hydrologic Engineering, 11(1), 71-79.

Rajaraman, A., & Ullman, J. D. 2012. Mining of Massive Datasets. Edinburgh UK: Cambridge University Press.

Schumacher, P., Olinsky, A., Quinn, J., & Smith, R. 2010. A comparison of logistic regression, neural networks, and classification trees predicting success of actuarial students. Journal of Education for Business, 85(5), 258-263.

Skillicorn, D. 2007. Understanding Complex Datasets: Data Mining with Matrix Decompositions. USA: Taylor & Francis Group.

Suh, S. C. 2012. Practical Applications of Data Mining: Jones & Bartlett Learning.

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology, 14(1), 137.

Witten, I. H., Frank, E., & Hall, M. A. 2011. Data Mining Practical Machine Learning Tools and Techniques (3rd Edition ed.): Morgan Kaufmann.