A study of extreme learning machine on small sample-sized classification problems

Boon Pin Ooi¹, Norasmadi Abdul Rahim¹, Maz Jamilah Masnan², and Ammar Zakaria³

¹Faculty of Electrical Engineering Technology, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
²Centre of Excellence for Social Innovation and Sustainability, Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
³Centre of Excellence for Advanced Sensor Technology (CEASTech), Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia

norasmadi@unimap.edu.my

Abstract. Extreme learning machine (ELM) is a special type of single hidden layer feedforward neural network that emphasizes training speed and optimal generalization. The ELM model proposes that the weights of hidden neurons need not be tuned, and the weights of output neurons can be calculated by finding the Moore-Penrose generalized inverse method. Thus, the ELM classifier is suitable to use in a homogeneous ensemble model due to the untuned random hidden weights which promote diversity even with the same training data. This paper studies the effectiveness of the ELM ensemble models in solving small sample-sized classification problems. The research involves two variants of the ensemble model: the normal ELM ensemble with majority voting (ELE), and the random subspace method (RS-ELM). To simulate the small sample cases, only 30% of the total data will be used as the training data. Experiment results show that the RS-ELM model can outperform a multi-layer perceptron (MLP) model under the assumptions of a Friedman test. Furthermore, the ELE model has similar performance as an MLP model under the same assumptions.

1. Introduction

With the advancement in computer technology, data collection becomes effortless and data storage becomes more affordable, a massive amount of data being generated due to this. The field of machine learning also benefits from the increasing amount of data collected, new techniques are being proposed on how to gain an advantage from this big data trend. However, there are situations when researchers will need to work on a small data set, information such as rare medical disease, startup product, as well as information with confidentiality are likely to be scarce. A data set is classified as small when the ratio of the number of samples and the number of features is less than 20 [1].

In the field of machine learning, a classification model is called as classifier. For a classifier to correctly classifies an input, the classifier is required to train on a set of training data. A neural network is a type of classifier that models the human brain cell that processes information through its interconnected weights. With the introduction of error back-propagation algorithms, the neural network becomes one of the most popular models in the pattern classification field. The back-
propagation algorithm is the barebone for some well-known neural network architectures, such as multi-layer perceptron, and convolution neural network [2]–[4].

In [5], the authors proposed a special type of neural network known as extreme learning machine (ELM). ELM is a single hidden layer neural network model that emphasizes training speed and optimal generalization. The model uses random weights for the hidden layer and Moore-Penrose generalized inverse method to calculate the output layer weights. This technique avoided the iteration approach by the back-propagation algorithm, thus greatly reduces the training time even for a large amount of data.

A major problem that small sample-sized (SSS) data classification suffers from is overfitting. Overfitting is the phenomenon that occurs when a classification model can achieve very high prediction accuracy with training data, but significantly lower when working with testing data. This happens due to the model has a high learning capacity, and with few samples to train from, the model attempts to “memorizes” all the training patterns, thus it can achieve very accurate prediction.

An approach to solve the overfitting issue is the classifier ensemble method [1], [6], [7]. The ensemble method proposed to combine two or more classifiers in a single classification task to further improve the accuracy. The type of ensemble model depends on the pool of base classifiers in use. A homogeneous ensemble model uses base classifiers from the same type, each base classifier will be trained on a different subset of data to induce diversity among the pool of base classifiers. On the other hand, the heterogeneous ensemble model includes different types of base classifiers in its pool.

The success of the ensemble model depends on the diversity included in the full model. In a Bootstrap Aggregating (Bagging) approach, diversity was introduced to the overall model by training each base classifier with a randomly selected subset from the training data [8]. The Adaptive Boosting (AdaBoost) technique assigns each training sample with a weight, the weight represents the probability of that sample being chosen into the next training subset [9]. This technique starts with a base classifier, then based on the performance of that classifier, increase the weight of incorrect samples so that those samples that incorrectly predicted have a higher chance to be selected in the next subset selection. Another popular ensemble method is the Random Subspace (RS) approach. The RS model selects feature subspace randomly to form a new training subset, this approach was proposed to solve the overfitting issue as stated in [10]. Due to the RS model creates new training subsets with reduced feature number, which indirectly increases the instance to feature ratio, thus is suitable to apply on SSS problems.

This research studies the performance of the RS ensemble model with ELM as the base classifier. This paper is structured as follows: Section 2 provides a brief background review of the literature. Section 3 explains the methodology used in this research. Section 4 presents the result of the experiment and discusses the observation. Finally, this paper ends with Section 6 by drawing a conclusion to this work.

2. Background review

This section reviews the basic of artificial neural network, ELM, and RS ensemble model.

2.1. Artificial neural network

An artificial neural network (ANN) is inspired by a human brain cell that processes information through synaptic weight. Given a data set with \( N \) number of samples and \( M \) number of features, a single neuron model can be expressed as equation (1), where \( y \) is the predicted output from the neuron, and \( x_m \) is the \( m \)-th feature from one sample. The weight that connects input \( x_m \) to the neuron is referred as \( w_m \), and \( b \) represents the bias at this neuron. Finally, the total sum of inputs multiplied by weights plus with the bias undergo an activation function, \( f \) to produce the predicted output of this neuron [11].
\[ y = f \left( b + \sum_{m=1}^{M} x_m w_m \right) \] (1)

A single neuron has very limited processing power, thus ANN is often referring as a network of neurons that separated into different layers as shown in figure 1, this architecture is known as multi-layer perceptron (MLP). The first layer in an MLP model is the input layer, this layer provides no processing power, it is a simple layer that serves data into the system. The last layer of the network is the output layer, this layer processes the output data from the previous layer and produces predicted output. The layers between the input layer and the output layer are the hidden layers, there can be more than one hidden layer in an MLP model.

![Multi-layer perceptron model with two hidden layers.](image)

**Figure 1.** Multi-layer perceptron model with two hidden layers. Hidden layer can have any number of neurons, and each neuron is fully interconnected to all neurons in the next layer.

### 2.2. Extreme learning machine

Extreme learning machine (ELM) is a special type of single hidden layer neural network. Most neural networks are trained with the error back-propagation algorithm, but the algorithm utilizes the gradient descend technique that employ s site rative tuning of weights. Due to that, the training of neural networks is usually slow, and could potentially falls into a local minimum [5], [12]. The ELM proposes that the weights of the hidden layer need not be tuned, and the weights of the output layer will be calculated by obtaining the Moore-Penrose generalized inverse as in equation (2). From the equation, \( T \) is the one-hot encoded target vector that serves as the ground truth of training data. The \( H^{\dagger} \) is the Moore-Penrose generalized inverse of hidden layer output, \( H \).

\[
\hat{\beta} = H^{\dagger} \cdot T
\] (2)

In [13], the authors proposed an ensemble based ELM (EN-ELM) to deal with the overfitting issue and improve predictive stability. The experiment used the cross-validation technique to split the data set into training and testing sets. All ELM base classifiers were trained on the full training set, as the weights of hidden neurons are randomly assigned, so it is safe to assume that no base classifiers are the same. VELM_AP proposed to prune the final pool of base classifiers through geometric mean
ranking, leaving only an odd number of base classifiers, and combine their result with majority voting [14]. The authors in [15] proposed F-ELM for the ELM classifiers to work with small sample data. The F-ELM utilized the Markov-Boundary approach to selects only the representative features, then deploy multiple ELM classifiers on the reduced set of data, finally combine the result with a weighted majority voting. This implies that by reducing the number of features in a data set, an ELM algorithm could perform better in SSS classification problems.

2.3. Random subspace
As it is suggested that there is no single best classifier, the ensemble model takes the advantage of multiple base classifiers and combines all results based on certain criteria [9]. The decisive factor to the success of an ensemble model relies on its base classifiers diversity [16]. In Bagging and AdaBoost, the diversity of the model is injected by training the base classifiers with different subsets of training samples. On the other hand, the random subspace (RS) approach proposed to split the overall feature space into several smaller subspaces [10].

In the case of the SSS problem, the approach to generate a training subset will cause the number of samples in each subset to even smaller, thus Bagging and AdaBoost are not suitable to use. While the RS method split the data in feature space, the number of samples in each subset will be the same as the number of samples in the overall training set [17]. Thus, the RS method can be used to reduce the number of features in high-dimensional data, and improve classification accuracy [18].

3. Methodology
This research proposes to solve the SSS classification problem by using a random subspace ensemble model with extreme learning machine as the base classifiers. The approach takes advantage of ELM fast training time to produce a pool of trained base classifiers, then combine results from all classifiers via a majority voting approach.

The experiment is conducted with 10 data sets, the properties of each data set are given as in table 1. The “contractions” and “laryngeal1” data sets are obtained in [19], and other data sets are retrieved from the UCI machine learning repository [20].

| Data set     | Sample | Feature | Class |
|--------------|--------|---------|-------|
| contractions | 98     | 27      | 2     |
| iris         | 150    | 4       | 3     |
| laryngeal1   | 213    | 16      | 2     |
| mfeat-fou    | 2000   | 76      | 10    |
| mfeat-kar    | 2000   | 64      | 10    |
| mfeat-pix    | 2000   | 240     | 10    |
| rds          | 85     | 17      | 2     |
| spect        | 267    | 22      | 2     |
| vowel        | 528    | 10      | 11    |
| weaning      | 302    | 17      | 2     |

For this research, four models will be investigated: a single extreme learning machine classifier (ELM), a multi-layer perceptron model (MLP) with a single hidden layer, a simple ensemble model with ELM (ELE), and a random subspace model with ELM (RS-ELM). For the ELE and RS-ELM, the ensemble size will be fixed at \( L = 25 \). The number of hidden layer neurons will be calculated with \( \sqrt{M \times C} \) where \( M \) is the number of features, \( C \) is the number of classes, and the result will be rounded to the nearest whole number. All ELM models will use radial basis activation function at the hidden neurons, and linear activation function for the output neurons.

The MLP model will be trained with the error back-propagation algorithm, the learning rate is fixed at 0.1. The training epoch is 100 with no early stopping condition. The sigmoid activation function is
used in the hidden neurons, and the output neurons use the softmax function. A negative log-
likelihood operates as the loss function to govern the weight tuning process.

This experiment will investigate the performance of each model over all data sets through the
Friedman test [21]. This part will run the experiment in 10 iterations, each iteration will split the data
set with the 3:7 hold-out method, 30% of the data will be used as the training set, and 70% of the data
will be used as the testing set. Each iteration will also carry out the training and testing phase 30 times.
Finally, all results from all data sets will be combined, and the Friedman test will be performed to
evaluate the ranking of each algorithm.

4. Discussion

Table 2 presents the sample-to-feature ratio for all 10 data sets. The ratio is calculated with the 30%
training set. Since a data set is considered as small when the sample-to-feature ratio is less than 20, all
data sets in this experiment maintain the SSS property, with the “vowel” data set has the highest ratio
of 15.8%.

To investigate the performance of the four models stated in Section 3, the experiment was carried
out with 10 iterations and each iteration will conduct 30 train-test runs, the results from table 3 are the
averaged accuracies over 300 runs. The bolded results represent the highest accuracy obtained over a
data set.

| Data set | ratio |
|----------|-------|
| contractions | 1.07 |
| iris | 11.25 |
| laryngeal | 4.06 |
| mfeat-fou | 7.89 |
| mfeat-kar | 9.38 |
| mfeat-pix | 2.50 |
| rds | 1.41 |
| spect | 3.68 |
| vowel | 15.80 |
| weaning | 5.41 |

| Data set | ELM | ELE | RS-ELM | MLP |
|----------|-----|-----|--------|-----|
| contractions | 54.75 | 61.46 | 71.10 | 80.93 |
| iris | 57.07 | 78.53 | 81.29 | 63.63 |
| laryngeal | 4.69 | 52.98 | 72.23 | 62.73 |
| mfeat-fou | 29.12 | 52.16 | 56.09 | 12.08 |
| mfeat-kar | 17.45 | 37.76 | 48.20 | 80.82 |
| mfeat-pix | 17.62 | 38.48 | 55.49 | 81.65 |
| rds | 47.64 | 46.24 | 46.24 | 49.18 |
| spect | 50.85 | 49.99 | 50.67 | 50.99 |
| vowel | 50.85 | 49.99 | 50.67 | 50.99 |

From table 3, we observed that the MLP model achieved the highest accuracies in half of the total
data sets, the RS-ELM model placed at second with 4 data sets, the ELE model has performed slightly
better than the RS-ELM model in the “vowel” data set, and the ELM model has the weakest
performance. Looking at the results from the “weaning” data set, the MLP model has a mean accuracy
of 50.99%, and the mean accuracy of the RS-ELM model is 50.67%, a slight difference of 0.32% is
not sufficient to prove that the MLP model is better than the RS-ELM model. Thus, for a proper
comparison, the Friedman test is conducted on the combined results from all data sets.

Figure 2 shows the critical difference graph drawn from the Friedman test result. The Friedman test
makes the comparison between multiple algorithms through a ranking system. The model with the best
performance in each comparison will be ranked at 1. Taking the mean of all ranking will decide the
final rank for that algorithm. In this experiment, the highest ranking is achieved by the RS-ELM model
with a mean rank of 1.85. The ELM model has the worst performance among all algorithms, ranked at
3.24. The MLP model and the ELE model ranked with 2.42 and 2.50, respectively.
Figure 2. The critical difference graph drawn from the Friedman test result. The red dashed-line indicates half CD from the mean rank of RS-ELM algorithm. Any “whisker” overlapping suggested that two algorithms are not significantly different.

The critical difference (CD) of this test is calculated as 0.086. As the Friedman test states that two algorithms are significantly different if and only if their mean rank is different by at least one CD. Thus, by comparing the mean rank of the RS-ELM model with the mean rank of the MLP model, the difference is 0.57, so it is confident enough to state that the RS-ELM model is significantly better than the MLP model. Finally, the mean rank difference between the MLP model and the ELE model is 0.08. The Friedman test claims that these two algorithms have no significant difference in performance.

5. Conclusion
This paper studies the effectiveness of the extreme learning machine (ELM) model in solving small sample-sized classification problems together with two ensemble variants: extreme learning machine ensemble (ELE) and random subspace with extreme learning machine (RS-ELM). The ELE model employs multiple ELM as base classifiers to train on the same training data, due to the random hidden weights of the ELM model, no data augmentation is needed to produce diversity among the pool of classifiers. The RS-ELM model reduces the overall feature space into a smaller subset, which indirectly increases the sample-to-feature ratio. The experiment results show that RS-ELM with an ensemble size of 25 outperforms all other algorithms in the experiment. Also, the ELE model has a similar performance as an MLP model trained with the error back-propagation algorithm under the analysis of a Friedman test. For future research, a regularized ELM with a self-pruning algorithm would better fit in the role as the base classifier since most overfitting issues were caused by an overly complex model. The ELM could start with a large number of hidden neurons, then prune out irrelevant neurons to obtain a “slimmer” network through the regularized self-pruning algorithm.

Acknowledgments
This research was supported by the Ministry of Higher Education, Malaysia, through the Fundamental Research Grant Scheme (Ref: FRGS/1/2020/ICT03/UNIMAP/02/2). The authors also acknowledge the Faculty of Electrical Engineering Technology, Universiti Malaysia Perlis (UniMAP) for the equipment and technical assistance.
References

[1] Ruparel N H, Shahane N M and Bhamare D P 2013 Learning from small data set to build classification model: A survey International Conference on Recent Trends in engineering & Technology pp 23–6

[2] Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors Nature 323 pp 533–6

[3] Alsmadi M, Omar K and Mohd Noah S A 2009 Back propagation algorithm: The best algorithm among the multi-layer perceptron algorithm IJCSNS Int. J. Comput. Sci. Netw. Secur. 9 pp 378–83

[4] Suliman A and Zhang Y 2015 A review on back-propagation neural networks in the application of remote sensing image classification J. Earth Sci. Eng. 5 pp 52–65

[5] Huang G B, Zhu Q Y, and Siew C K 2004 Extreme learning machine: A new learning scheme of feedforward neural networks IEEE International Joint Conference on Neural Networks pp 985–90

[6] Chawla N, Moore T E, Bowyer K W, Hall L O, Springer C and Kegelmeyer P 2001 Bagging is a small-data-set phenomenon IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2 pp 684–9

[7] Zaman M F and Hirose H 1996 Classification performance of bagging and boosting type ensemble methods with small training sets New Gener. Comput. 29 pp 277–92

[8] Breiman L 1996 Bagging predictors Mach. Learn. 24 pp 123–40

[9] Dietterich T G 2000 Ensemble methods in machine learning Multiple Classifier Systems 1857 pp 1–15

[10] Ho T K 1998 The random subspace method for constructing decision forests IEEE Trans. Pattern Anal. Mach. Intell. 20 pp 832–44

[11] Prieto A, Prieto B, Ortigosa E M, Ros E, Pelayo F, Ortega J and Rojas I 2016 Neural networks: An overview of early research, current frameworks and new challenges Neurocomputing 214 pp 242–68

[12] Huang G B, Zhu Q Y and Siew C K 2006 Extreme learning machine: Theory and applications Neurocomputing 70 pp 489–501

[13] Liu N and Wang H 2010 Ensemble based extreme learning machine IEEE Signal Process. Lett. 17 pp 754–7

[14] Shukla S and Yadav R N 2015 Voting based extreme learning machine with accuracy based ensemble pruning Int. J. Comput. Appl. 115 pp 14–8

[15] Yin Y, Zhao Y, Li M and Zhang B 2016 An enhanced extreme learning machine for efficient small sample classification Proc. ELM-2015 1 pp 501–9

[16] Dietterich T G 2000 An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization Mach. Learn. 40 pp 139–57

[17] Kotsiantis S 2011 Combining bagging, boosting, rotation forest and random subspace methods Artif. Intell. Rev. 35 pp 223–40

[18] Li X and Zhao H 2009 Weighted random subspace method for high dimensional data classification Stat Interface 2 pp 153–9

[19] Kuncheva L I 2013 A bound on kappa-error diagrams for analysis of classifier ensembles IEEE Trans. Knowl. Data Eng. 25 pp 494–501

[20] Bache K and Lichman M 2013 UCI machine learning repository University of California Irvine School of Information (http://www.ics.uci.edu/~mlearn/MLRepository.html)

[21] Demšar J 2006 Statistical comparisons of classifiers over multiple data sets J. Mach. Learn. Res. 7 pp 1–30