Metaheuristic Training Neural Networks

April 20th, 2021

Abstract

This paper explores the use of metaheuristics in training a feed-forward artificial neural network, examining each algorithm and their performances in training a neural network versus backpropagation. Results show promise for the use of certain metaheuristics although the go-to approach of backpropagation is still superior in many ways.

1 Introduction

2 Metaheuristics

Metaheuristics are a type of "informed" local-global search which aim to optimize an objective function [1]. If the search space of a problem is known, it may be an appropriate scenario where a metaheuristic can optimize this problem. These algorithms are often accompanied by or designed with a metaphor in mind, which has lead to controversy [8].

2.1 Genetic Algorithm

Genetic algorithm (**GA**) takes cues from evolutionary biology to evolve solutions to a problem based on both population and individual agent traits [3].

2.2 Particle Swarm Optimization

Particle swarm optimization (**PSO**) is inspired by the flocking psychology of birds or other swarming animals [5].

2.3 Differential Evolution

Not too dissimilar to GA, differential evolution (**DE**) is another evolution-based metaheuristic but instead of inspired by biology, evolves solutions based on arithmetic and combining solutions according to some mathematical rules [7].

For DE, a population of solutions is generated randomly and are moved through the search space using arithmetic rules which allow combination of solutions. DE uses two parameters: *crossover rate*, which determines the frequency in which solutions are combined, and *differential weight*, which is a multiplicand determining how much of the difference between two crossovered solutions is imparted to the new solution.

While GA considers two parent chromosomes to generate two child chromosomes, DE contrastingly considers three "parent" solutions, combined in some ratio, to form a single new "child" solution. DE does not offer similar selection strategies like elitism as in GA, and instead a new population is guaranteed to be unique between generations.

Solutions evolve like so: for every solution x in the population, three other solutions are chosen, a, b, c such that x, a, b, c are distinct entities from each other. A solution y is generated according to the below:

$$y_i = \begin{cases} a_i + F \times (b_i - c_i) & \text{if } r_i < CR \text{ or } i = R \\ x_i & \text{otherwise} \end{cases}$$

y is generated in sequence of axes $i = 1, 2, \ldots, n$. For every solution that is created this way, a random index $R \in \{1, 2, \ldots, n\}$ is chosen which signifies an axis which is always updated, even if crossover is not to occur. For this equation, CR is the crossover rate and F is the differential weight.

Once y is found, it is compared to x, and if it has a better fitness, then x is replaced with y in the population and the process continues for other agents.

This type of evolution is not strictly bio-inspired and instead is more mathematical in nature. Crossover for DE allows the metaheuristic to have an *exploratory factor*, where the search space is explored in larger steps when compared to an *exploitative factor*, where solutions are fine-tuned in place to reach optima around the current solution position. DE still exhibits an exploitative factor, as when the population converges to an optima, all solutions within the population more closer resemble each other, meaning exploration is minute.

Over time, the best fitness of solutions in the population is improved until convergence (either to a global optimum or local optima). Every generation is assured to either remain the same or improve since a fitness comparison is the criteria for a solution being replaced. In other words, solutions will never regress to a worse solution.

2.4 Bat Algorithm

Bat algorithm (**BA**) is similar to other metaheuristics in that it is a swarm-based approach to improving solutions. Solutions within the search space are updated similar to how swarming microbats descend onto prey or food [9].

3 Network Training

A neural network is bio-inspired tool under the broader class of machine learning algorithms. It consists of a sequence of layers, themselves containing nodes—called neurons. The biology analogy is the brain, where this technique of machine learning takes its namesake [6]. Neural networks can be used for many tasks, although the classical usage is one of classification: using a neural network to train a model which can be used as a classifier.

The layers between neural network layers are composed of synaptic weights: these weights are mathematical links between neurons and signify how strong a signal between neurons is. In the simplest sense, a synaptic weight is some multiplicand between neurons, altering the signal between them.

One of the most popular approaches to training a neural network is using *backpropagation* (**BP**), which is a gradient-based training mechanism which relies on the existence of an error gradient between a target output value and the actual found output value.

Training patterns are inputted into the network at the first layer, travel to subsequent layers through a *feedforward* process, then when in the output layer, the error is calculated as a function of target and actual values. The error signal is backpropagated through the network in reverse, adjusting the synaptic weights as needed so the model more closely resembles the training pattern.

Given enough training patterns and enough iterations or repetitions of the process, ultimately—and provided appropriate network parameters are chosen—the network model is able to classify not only previous data fed into it, but also novel data from the same data set.

However, BP is mathematically robust yet conceptually obtuse: oftentimes neural networks are referred as black-boxes, as while the inputs and outputs are known, the inner workings are abstract. Thus while BP is an adequate training mechanism for a neural network, alternatives exist such as the use of metaheuristics.

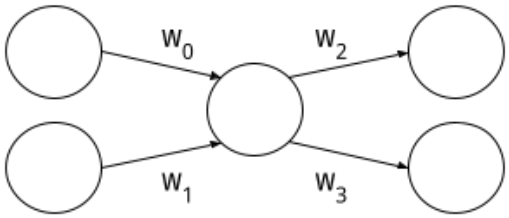

Figure 1: Connections between layers can be enumerated as a set of weights.

In the previous figure, a simple 2-1-2 network is shown and each synaptic weight connecting each layer is enumerated. For this network, which presents a 4-dimensional problem space, it has four weights.

Since metaheuristics explore the search space, it's useful to think of solutions as a position in n-space, where n is the dimensionality of the problem. For this example network, solutions update themselves in a 4-dimensional search space.

Each metaheuristic tested either takes literal sense of this concept—bat algorithm having bat agents with a specific position, as example—or are more abstract or have no close conceptual similarity to an agent "moving". Nevertheless, they all function in an identical fashion.

Since synaptic weights are enumerable in this way, it is simple to decode the network as instead a position in n-space. For this reason, a network can be be represented as a solution for a metaheuristic.

To use an example, a GA chromosome can be created as some encoding of the network synaptic weights. A network of weights $[w_0, w_1, ..., w_n]$ can be encoded as a chromosome of genes $[x_0, x_1, ..., x_n]$, where $x_i = w_i$. Likewise, this encoding is reversible, where for this sample a chromosome can be translated into a network.

This translation is integral to the inter-connectivity between a network and its metaheuristic training mechanism, as a metaheuristic cannot be perform on a network directly.

In the previous figure it is shown how a metaheuristic is integrated within a neural network's training process. For this figure, d is the dimensionality of the network—how

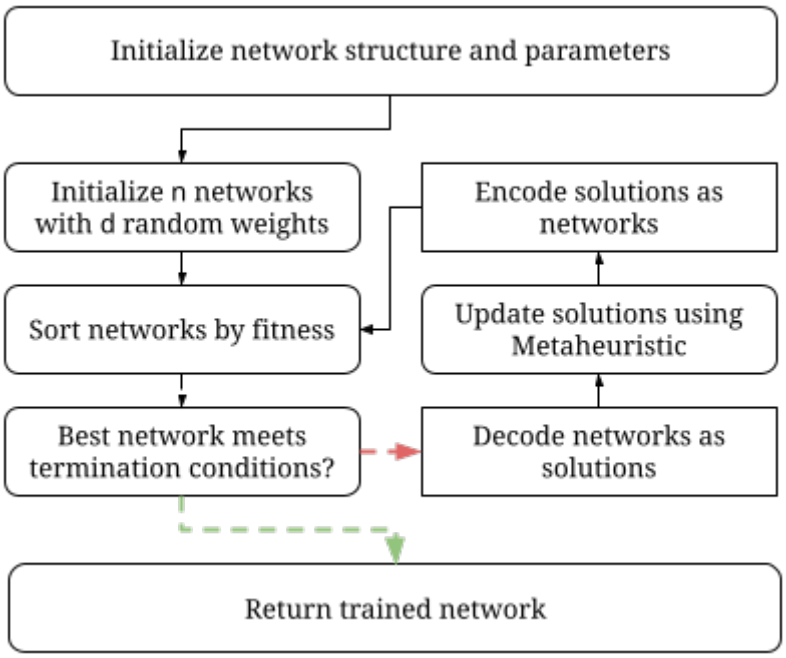

Figure 2: Metaheuristic adapted for training a neural network.

many synaptic weights exist in the network—and n is the number of agents for use with the metaheuristic.

Since metaheuristics are population based, this method of training neural networks is analogous to training laterally multiple networks at once and using both the traits of the population and individual solutions to improve.

4 Experimental Setup

To evaluate the efficacy of metaheuristics in training a neural network, a standard feed-forward network was implemented alongside implementation of each training mechanism: backpropagation (**BP-NN**), genetic algorithm (**GA-NN**), particle swarm optimization (**PSO-NN**), differential evolution (**DE-NN**), and bat algorithm (**BA-NN**).

Further, a testing suite was devised to train a model using each training mechanism and collect the results in the form of error over time curves. The error function used is mean squared error (**MSE**).

This suite performs each training 100 times for each data set and then discards the worst 20% of runs, as the mean of runs should be indicative of generally good results. A poor training attempt would normally be discarded and here is no different. Runs are then aggregated and the mean is found to be plotted.

Training ceases when one of two conditions are met: either the MSE of the network has reached a value of 0.1 or below, or the maximum number of epochs has elapsed. In the case where convergence did not occur without the defined number of epochs, that training technique is determined to have not converged.

Consideration has to be given for network and training parameters which vary by data set. Additionally, the data sets themselves need to be examined.

4.1 Data

Six data sets were used: *Iris* [2] for classification of species of iris flowers, *Penguins* [4] for classification of penguin species, *Wheat Seeds* [2] for classification of wheat varieties, *Wine* [2] for classification of wine varieties, *Breast Cancer* [2] for classification of breast cancer diagnosis, and *Ionosphere Radar* [2]. These data sets were chosen as they increase in complexity at a generally consistent rate, allowing for scalability of metaheuristic training to be considered.

Rows in the data set comprise of input patterns which are separated into attributes—of which there are many—and classifications—of which there is one, albeit with multiple possible values. The network structured initialized for a given data set is a consequence of these variables: the input layer contains a number of neurons equal to the number of attributes of the data, and the output layer contains a number of neurons equal to the number of distinct possible classifications of the data.

Data Set	**Instances**	**Attributes**	**Classes**
Iris	151	4	3
Penguins	333	6	3
Wheat Seeds	211	7	3
Wine	178	13	3
Breast Cancer	570	31	2
Ionosphere Radar	352	34	2

Table 1: Data sets used for results collection.

In the previous table, this information is tabulated. While the input and output layers of the network is decided by the data, the hidden layer size is arbitrary but still chosen through some experimentation. A smaller hidden layer will train faster, but too small will result in poor results. Experimentation was performed to minimize the hidden layer size while still maintaining satisfactory results.

Data Set	**IL Size**	**HL Size**	**OL Size**	**Dimensionality**
Iris	4	3	3	27
Penguins	6	4	3	43
Wheat Seeds	7	5	3	58
Wine	13	6	3	105
Breast Cancer	31	8	2	274
Ionosphere Radar	34	10	2	372

Table 2: Dimensionality of each problem.

In the previous table, the dimensionality of each classification problem is found as a function of network topology. Declaratively, the dimensionality of a network problem is found using the below equation:

$$d = (HL \times (IL + 1)) + (OL \times (HL + 1))$$

Dimensionality is related to how "difficult" the classification problem is: while data can be intrinsically complex, dimensionality gives a general idea of how complicated it will be to train a network using that data.

4.2 Parameters

Each network is initialized with a few parameters, some of which are specific to each metaheuristic used. Many oarameters remain consistent between training methods, but many parameters are initialized on a per-problem basis.

All parameters are chosen through some trial-and-error as there are often no "one size fits all" approach to parameter selection and tuning. Parameters chosen provide generally good results but could be improved, perhaps by an optimization algorithm.

4.2.1 Consistent Parameters

Parameters related to network topology and the training process are consistent between data sets and metaheuristics. Three parameters, the number of epochs, the holdout ratio, and the initial weight range are fixed for all problems.

Epochs	**Holdout Ratio**	**Initial Weight Range**
100	0.70	[-0.50, 0.50]

Table 3: Constant network parameters.

In the previous table, the three constant network parameters are tabulated. Epochs is the amount of training iterations performed and is one of the termination conditions for training, the holdout ratio is the proportion of input data to be used for training the network versus testing the network, and the initial weight range is the range between minimum and maximum values for weight initialization.

Since this is a comparative study between BP and metaheuristics, BP as the default decides many of these parameters. For example, for a metaheuristic to be relatively viable, the number of epochs should not be too dissimilar for metaheuristics: to this end, 100 epochs was chosen as BP could train within this amount in every scenario. The same applies for the other two parameters.

Likewise, hidden layer size is a constant, but only between training methods: this parameter varies only by data set used.

Data Set	**Hidden Layer Size**
Iris	3
Penguins	4
Wheat Seeds	5
Wine	6
Breast Cancer	8
Ionosphere Radar	10

Table 4: Hidden layer size per data set.

In the previous table, the hidden layer size is tabulated per data set. Since the difference between BP and metaheuristics is being examined, this needs to remain equal as all training methods are effectively training identical networks.

4.2.2 BP-NN

BP relies on two parameters: *learning rate*, or the amount of correction a synaptic weight should take from the training instance, and *momentum rate*, which considers some portion

of the prior training instance when updating using the current training instance.

Data Set	LR	MR
Iris	0.100	0.001
Penguins	0.100	0.001
Wheat Seeds	0.100	0.001
Wine	0.100	0.002
Breast Cancer	0.100	0.003
Ionosphere Radar	0.100	0.002

Table 5: Backpropagation parameters.

In the previous table, BP-NN parameters are tabulated per data set, where LR, MR are learning rate and momentum rate respectively.

Since BP is a single-agent training mechanism, it does not have a population as metaheuristics do.

4.2.3 GA-NN

For the first metaheuristic, GA considers six parameters for training. They are the population size (how many chromosomes), crossover rate, mutation rate, elite proportion, tournament proportion, and a base mutation value.

Data Set	Population	CR	MR	E_p	T_p	Base
Iris	100	0.90	0.03	0.05	0.03	0.5
Penguins	100	0.90	0.03	0.05	0.03	0.5
Wheat Seeds	100	0.90	0.04	0.05	0.04	0.6
Wine	100	0.90	0.05	0.05	0.05	0.7
Breast Cancer	100	0.90	0.05	0.05	0.07	0.8
Ionosphere Radar	100	0.90	0.06	0.05	0.09	0.9

Table 6: Genetic algorithm parameters.

In the previous table, GA-NN parameters are tabulated per data set, where CR, MR, E_p, T_p are the crossover rates, mutation rates, elite proportions, and tournament proportions respectively.

4.2.4 PSO-NN

PSO has five different parameters controlling its movement: population size, inertial weight, cognitive and social coefficients, and a boundary variable.

In the previous table, these parameters are tabulated per data set, where ω is the inertial weight and c_1, c_2 are the cognitive and social coefficients respectively.

4.2.5 DE-NN

DE only has three parameters: population size, crossover rate, and differential weight.

Above are the DE-NN parameters tabulated, where CR is the crossover rate and F is the differential weight.

Data Set	Population	ω	c_1	c_2	Boundary
Iris	100	0.5	1.5	1.2	3
Penguins	100	0.5	1.4	1.3	4
Wheat Seeds	100	0.6	1.3	1.1	5
Wine	100	0.3	1.6	1.4	7
Breast Cancer	100	0.4	1.4	1.1	7
Ionosphere Radar	100	0.3	1.3	1.3	9

Table 7: Particle swarm optimization parameters.

Data Set	Population	CR	F
Iris	50	0.90	0.25
Penguins	50	0.90	0.35
Wheat Seeds	50	0.90	0.25
Wine	50	0.90	0.20
Breast Cancer	50	0.90	0.15
Ionosphere Radar	50	0.90	0.10

Table 8: Differential evolution parameters.

4.2.6 BA-NN

Lastly, BA has six parameters: population size, frequency minimum and maximum, a boundary, a loudness decreasing factor, and a pulse rate increasing factor.

Data Set	Population	F_{min}	F_{max}	Boundary	α	γ
Iris	100	0	2	3	0.90	0.90
Penguins	100	0	2	4	0.90	0.90
Wheat Seeds	100	0	2	5	0.90	0.90
Wine	100	0	2	7	0.90	0.90
Breast Cancer	100	0	2	7	0.90	0.90
Ionosphere Radar	100	0	2	9	0.90	0.90

Table 9: Bat algorithm parameters.

Here tabulated, the BA-NN parameters F_{min}, F_{max} are the frequency minimum and maximum values and α, γ are the loudness decreasing factor and pulse rate increasing factor respectively.

5 Experimental Results

Every network is trained according to the previously defined test structures. Plots are then generated allowing a visual aid to compare. Two statistical tests are also performed on the data: an ANOVA test and Tukey's HSD test.

The ANOVA test is designed to determine if multiple samples come from the same population; in other words, the ANOVA test is used for testing purposes to identify if all training techniques are the same or there is non-trivial differences between them. An ANOVA test produces two statistics that are relevant: an F-score, where an $F >=$

1.0 denotes confidence in the test, and a p-value, where a $p < 0.05$ suggests there is significance in choosing some training techniques over the others.

Tukey's HSD test, if an ANOVA test shows significance, will then identify where outliers exist by doing pairwise comparison of means and variances between samples. If better training methods exist, they will become evident more directly using this test.

Plots show a curve for each training method, where each curve is the MSE over time. At the bottom of each plot is a tickmark which denotes at what epoch, on average, the training method reached a termination condition.

5.1 Iris

Using the Iris data set and network parameters as defined previously for each training method, the below plot is found:

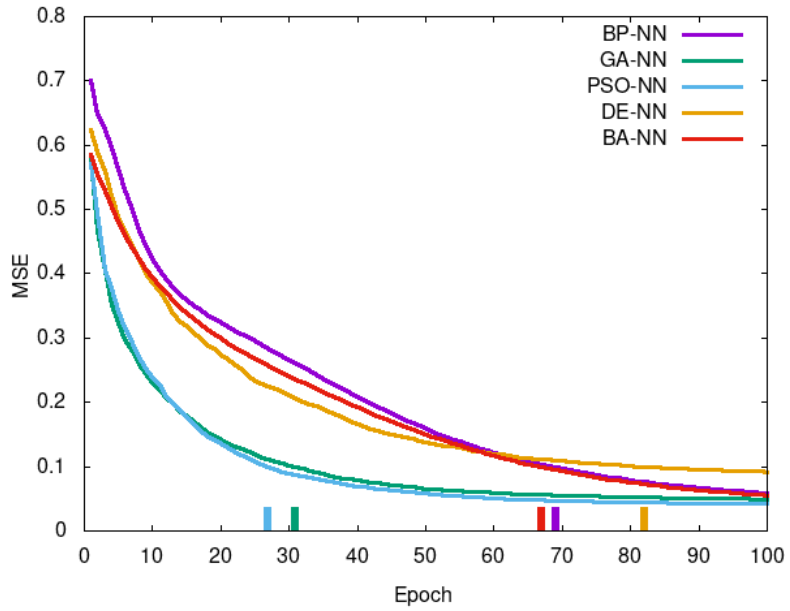

Figure 3: Iris test results for metaheuristics and backpropagation.

Visually, two training techniques (PSO-NN and GA-NN) converged to a termination condition much faster than the other three.

To determine if this difference is of statistic importance, an ANOVA test is run using each curve as a sample. This test yields an $F = 17.65$ and a $p < 0.001$ which suggests there is high confidence that there are some training methods that outperform the others.

Tukey's HSD test is then performed showing no significance between GA-NN and PSO-NN. Likewise it shows no significance between BP-NN, DE-NN, and BA-NN. However, samples between each grouping does show significance.

Since Tukey's HSD test showed significance between two groups, on comparing the mean number of epochs it took to reach termination condition, it can be concluded that GA-NN and PSO-NN outperform the other training methods and to a largely significant degree.

	BP-NN	**GA-NN**	**PSO-NN**	**DE-NN**	**BA-NN**
Min	51	9	10	36	31
Avg	70	31	27	82	67

Table 10: Minimum and average time to terminate for Iris data set.

Tabulated above is the minimum and mean number of epochs it took to reach a termination condition, green denoting best, and red worst. Here, BP-NN was the slowest network to converge, but on average, DE-NN was the slowest to converge. Agreeing with our results from before, both GA-NN and PSO-NN outperform the others significantly, having the fastest training at once and on average.

5.2 Penguins

The Penguins data set is then trained upon using network parameters as defined previously for each training method, producing the below plot:

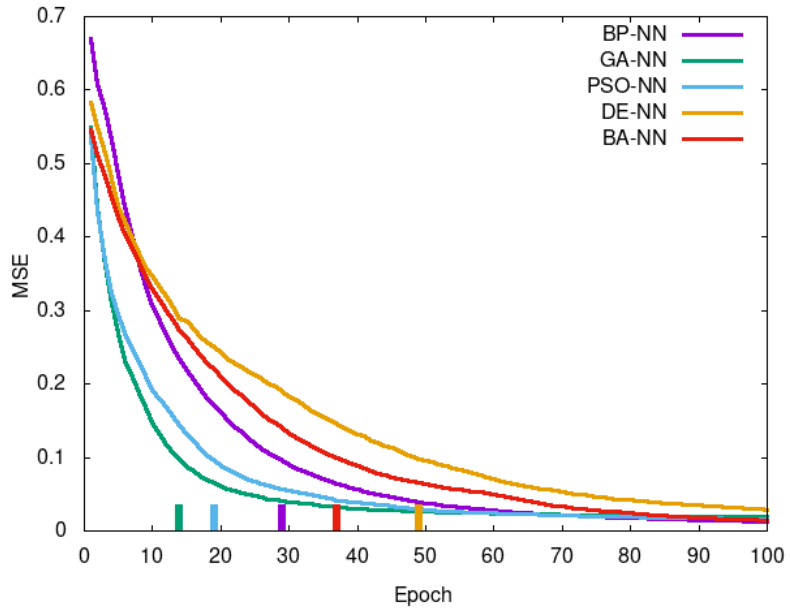

Figure 4: Penguins test results for metaheuristics and backpropagation.

It is difficult to discern visually whether curves come from the same distribution, so again an ANOVA test is performed. This test yields an $F = 9.0332$ and $p < 0.001$ which suggests there is confidence in curves not coming from the same population: outliers exist between training methods.

To identify them, Tukey's HSD test is performed, showing significance only between DE-NN and both GA-NN and PSO-NN. Here DE-NN is an outlier as the other samples show no significant difference.

From these tests, it is shown there is no significance in choosing between GA-NN, PSO-NN, BP-NN, or BA-NN; however, DE-NN is significantly slower than the other four.

	BP-NN	**GA-NN**	**PSO-NN**	**DE-NN**	**BA-NN**
Min	20	7	6	26	18
Avg	29	14	19	49	37

Table 11: Minimum and average time to terminate for Penguins data set.

In the previous table, as before the minimum and average number of epochs to reach a termination condition is given for each network type. As well, as before both GA-NN and PSO-NN are the fastest to train either as a single training or on average. DE-NN as found using the tests before was the slowest in every regard.

5.3 Wheat Seeds

Then, the Wheat Seeds data set is trained upon to attempt to make a model at classifying the data. Using the defined network parameters, the below plot is found:

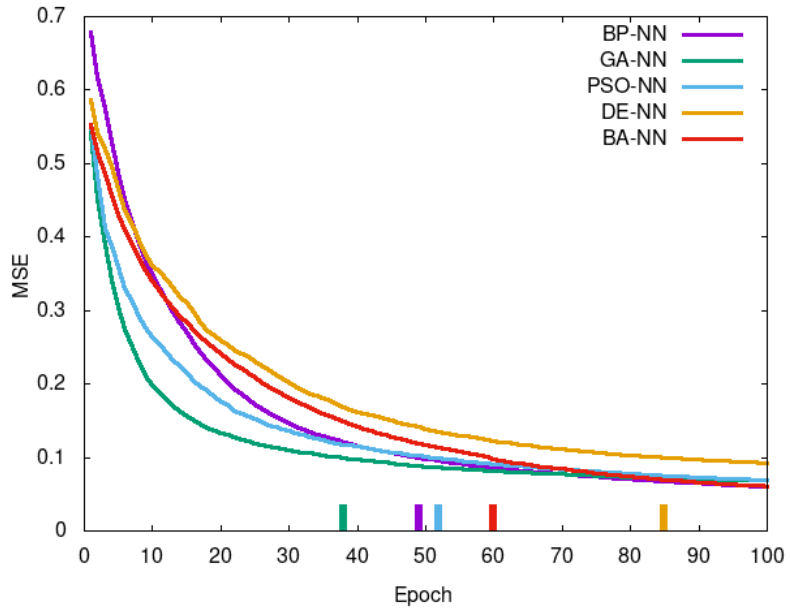

Figure 5: Wheat Seeds test results for metaheuristics and backpropagation.

To discover whether the curves are different to any meaningful degree, an ANOVA test is performed. This test yields an $F = 5.847$ with a $p < 0.001$, showing confidence that there is strong stochastic dominance between one or more samples.

Using Tukey's HSD test, each sample is examined and it is found that all samples—except DE-NN—come from the same distribution. This suggests there is no meaningful difference in choosing BP-NN, GA-NN, PSO-NN, or BP-NN. The only outlier is DE-NN.

While the ANOVA test showed significance, the significance is removed were DE-NN not a candidate sample for the tests, as pairwise difference of means between the other samples is insignificant.

	BP-NN	**GA-NN**	**PSO-NN**	**DE-NN**	**BA-NN**
Min	29	15	16	53	27
Avg	49	38	52	85	60

Table 12: Minimum and average time to terminate for Wheat Seeds data set.

Upon looking at the table above, this is the first test where GA-NN was fastest to converge in single-run training and on average. For the smaller problems, GA-NN was best in one or the other but it is interesting to see it outperform all other methods tested. However, since there is no significance in choosing it over the others, these results are inconclusive.

5.4 Wine

The Wine data set is trained upon using the previously defined network parameters. The below plot is made using those results:

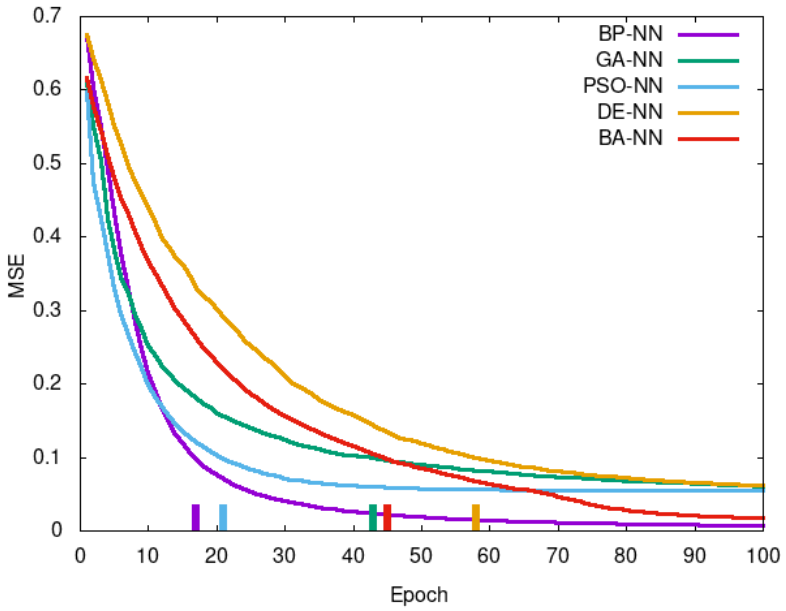

Figure 6: Wine test results for metaheuristics and backpropagation.

It is hard to determine which samples come from the same distribution, so an ANOVA test is performed yielding an $F = 12.081$, suggesting strong confidence in the test, and a $p < 0.001$ meaning strong significance in choosing some training methods over others.

To determine this, Tukey's HSD test is performed and the results are examined. The test suggests the curves come from three distributions: BP-NN and PSO-NN belong to the same distribution with strong confidence; GA-NN and BA-NN likewise belong to the same distribution; DE-NN is an outlier to each.

From this test, it can be surmised there is strong favor to choosing BP-NN or PSO-NN over other training methods.

	BP-NN	**GA-NN**	**PSO-NN**	**DE-NN**	**BA-NN**
Min	14	19	13	37	19
Avg	17	43	21	58	45

Table 13: Minimum and average time to terminate for Wine data set.

As before, a table is generated of the minimum and average number of epochs it took for each training method to converge.

Since BP-NN and PSO-NN have the smallest mean and minimum number of epochs to converge and they belong to the same distribution, either training method is preferable to the others. Of methods to choose, however, DE-NN is the outlier with the worst performance.

Interestingly, up to this point, GA-NN has been fastest to converge in single runs, on average, or both. This is the first problem where it is neither.

5.5 Breast Cancer

The Breast Cancer data set is then trained upon which yields the below plot using the defined network parameters for each training method:

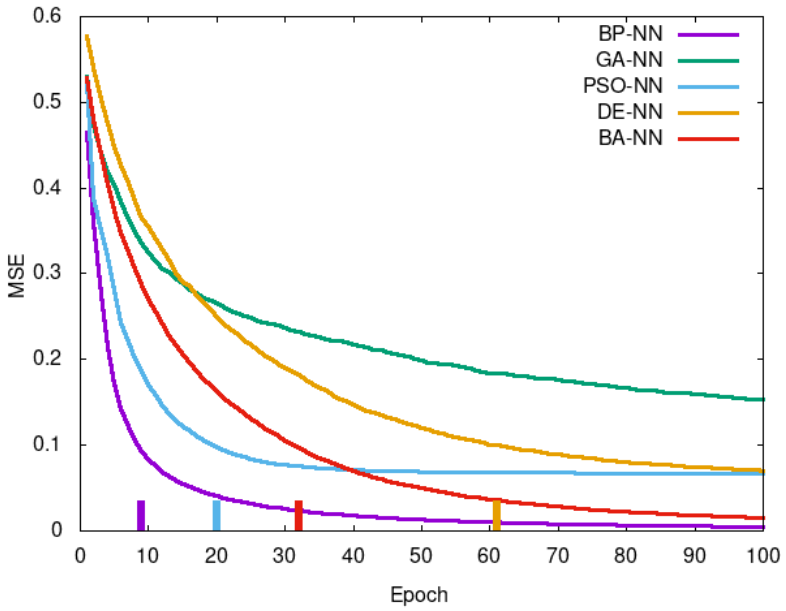

Figure 7: Breast Cancer test results for metaheuristics and backpropagation.

As with the Penguins data set, it is difficult to discern how related or unrelated each curve is other than GA-NN not converging.

To understand the plot better, an ANOVA test is performed yielding an $F = 61.494$ and $p < 0.001$ suggesting strong significance in choosing one or some training methods over others.

Tukey's HSD test is again performed to determine which samples come from the same distribution. The test reveals strong favor in choosing BP-NN over other methods. There is no significance in choosing PSO-NN over BA-NN or vice-versa, but there is slight significance in choosing BP-NN other either.

	BP-NN	**GA-NN**	**PSO-NN**	**DE-NN**	**BA-NN**
Min	6	52	11	37	16
Avg	9	DNC	20	61	32

Table 14: Minimum and average time to terminate for Breast Cancer set.

In the table above, performance is measured for each training method by way of the mean and minimum amount of epochs to reach termination condition. A "DNC" here, under GA-NN, means the training method did not converge.

As suggested by Tukey's HSD test, BP-NN has the speed advantage over the other training methods for the Breast Cancer classification problem.

5.6 Ionosphere Radar

The final data set, Ionosphere Radar, is test and plotted according to the previously defined network parameters:

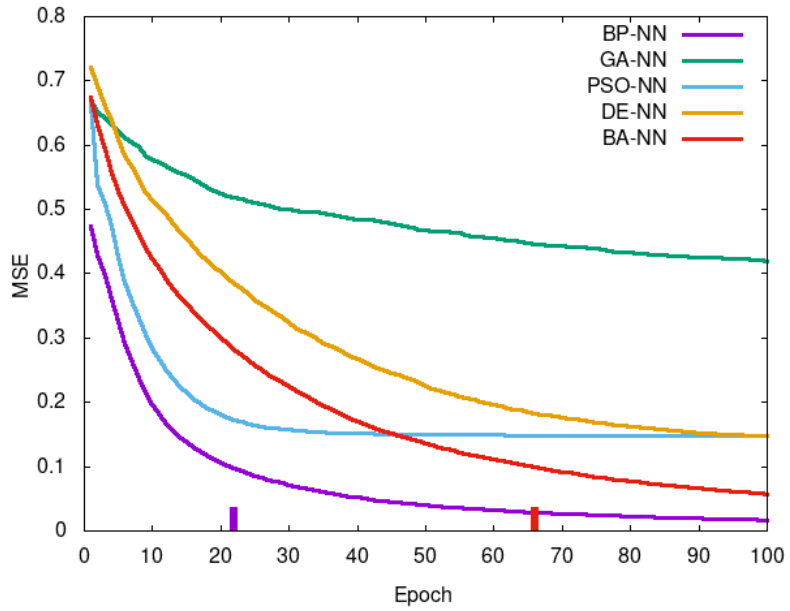

Figure 8: Ionosphere Radar test results for metaheuristics and backpropagation.

Here, only two training techniques converged within the allotted number of training iterations: BP-NN and BA-NN.

Visually, it is clear there is difference to the two methods, but to verify, an ANOVA test is performed, producing an $F = 185.7$, a very high F-score, and a $p < 0.001$, concluding that there is a large degree of significance between BP-NN and BA-NN.

Tukey's HSD test shows pairwise significance between samples as significant for all samples. What can be deduced from this is that between the two methods that converged, BP-NN is considerably faster.

	BP-NN	**GA-NN**	**PSO-NN**	**DE-NN**	**BA-NN**
Min	9	DNC	24	84	23
Avg	22	DNC	DNC	DNC	66

Table 15: Minimum and average time to terminate for Ionosphere Radar set.

Finally, a table is constructed showing the minimum and average number of epochs to reach a termination condition. As before, a "DNC" signifies the training method did not converge.

While PSO-NN and DE-NN could converge, on average they were unable to. Notable is GA-NN was unable to converge during any training runs. As alluded to from the ANOVA and Tukey HSD tests, BP-NN is the fastest training method for this problem.

6 Results Summary

7 Conclusion

References

[1] L. Bianchi, M. Dorigo, L. M. Gambardella, and W. J. Gutjahr. A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing: An International Journal*, 8(2):239–287, June 2009.

[2] D. Dua and C. Graff. UCI machine learning repository, 2017.

[3] D. Goldberg, G. David Edward, D. Goldberg, and V. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Artificial Intelligence. Addison-Wesley Publishing Company, 1989.

[4] A. M. Horst, A. P. Hill, and K. B. Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. R package version 0.1.0.

[5] J. Kennedy and R. C. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 1942–1948, 1995.

[6] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[7] R. Storn and K. Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. of Global Optimization*, 11(4):341–359, Dec. 1997.

[8] K. Sörensen. Metaheuristics—the metaphor exposed. *International Transactions in Operational Research*, 22(1):3–18, 2015.

[9] X.-S. Yang. *A New Metaheuristic Bat-Inspired Algorithm*, pages 65–74. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.