

Evolutionary Ensemble for *In Silico* Prediction of Ames Test Mutagenicity

Huanhuan Chen and Xin Yao

The Centre of Excellence for Research in Computational Intelligence and Applications
School of Computer Science, University of Birmingham, UK

Abstract. Driven by new regulations and animal welfare, the need to develop *in silico* models has increased recently as alternative approaches to safety assessment of chemicals without animal testing. This paper describes a novel machine learning ensemble approach to building an *in silico* model for the prediction of the Ames test mutagenicity, one of a battery of the most commonly used experimental *in vitro* and *in vivo* genotoxicity tests for safety evaluation of chemicals. Evolutionary random neural ensemble with negative correlation learning (ERNE) [1] was developed based on neural networks and evolutionary algorithms. ERNE combines the method of bootstrap sampling on training data with the method of random subspace feature selection to ensure diversity in creating individuals within an initial ensemble. Furthermore, while evolving individuals within the ensemble, it makes use of the negative correlation learning, enabling individual NNs to be trained as accurate as possible while still manage to maintain them as diverse as possible. Therefore, the resulting individuals in the final ensemble are capable of cooperating collectively to achieve better generalization of prediction. The empirical experiment suggest that ERNE is an effective ensemble approach for predicting the Ames test mutagenicity of chemicals.

Keywords: Ames Test Mutagenicity, *In silico* models, Evolutionary Ensemble, Negative Correlation Learning.

1 Introduction

The Ames test in *Salmonella typhimurium* is an *in vitro* biological assay to assess the mutagenic potential of chemical compounds. It is also considered as a quick assay to estimate the carcinogenic potential of a compound. Hence it serves as one of a battery of the most commonly used experimental *in vitro* and *in vivo* genotoxicity tests for safety evaluation of chemicals. Nevertheless, driven by new regulations and animal welfare, recently the needs of development of *in silico* models as alternative approaches to mutagenicity assessment of chemicals without animal testing is constantly increasing, and has attracted much attention from researchers in both fields of toxicology and computer science. Machine learning technique inevitably plays a major role in establishing relationships between chemical structural descriptors and mutagenicity for reducing, refining or replacing (3R) animal testing.

The classification problem addressed in this paper is to predict whether an Ames test of a chemical is positive or negative. A positive test outcome means that the chemical tested is more likely to be a mutagen whilst a negative test outcome means that the chemical is more likely a non-mutagen. Application of different classification methods to the prediction of Ames test mutagenicity has been studied already. While most of studies use single learner based classification methods such as decision trees, k -nearest neighbors, neural networks [2,3], and support vector machine [4], few use ensemble methods such as bagging, boosting, and random forest [5]. However, ensemble approaches have generally been shown to be superior to their corresponding base classifiers for most of the classification problems in machine learning [6]. The preliminary analysis using a variety of classifiers also showed us similar scenarios. First, ensemble approaches of bagging, boosting and random forests using the decision tree as a base classifier outperform non-ensemble classifiers. Second, random forest approach performs best among three ensemble approaches [7]. The superiority of random forest approach is largely due to that it uniquely adopts the feature subset selection, which is particularly of value for handling a data set with a large amount of feature variables [8]. Nevertheless, randomization of both data and features taken in random forests seems to pay more attention to diversity than to accuracy of individual classifiers in an ensemble, which may degrade the performance of the generated ensemble. Both theoretical [9] and empirical studies [10] demonstrated that the generalization ability of ensemble depends crucially on both accuracy and diversity among individual classifiers in the ensemble. In this study, we focus on a novel machine learning ensemble approach based on NNs and evolutionary computation, which has the potential to pay attention to both diversity and accuracy of individuals within an ensemble. We are particularly interested in an investigation of its efficacy of the approach to building *in silico* models for the prediction of the Ames test mutagenicity of chemicals.

It is non-trivial to design an accurate yet diverse ensemble due to a trade-off between accuracy and diversity in the ensemble. In [1], we have proposed to incorporate evolutionary random neural network ensemble with negative correlation learning [11] (ERNE) to design accurate and diverse ensemble for machine learning problems. Experimental results of ERNE have shown improvement over the existing ensemble algorithms, i.e. Bagging, Adaboost and random forests. Since ERNE employs bootstrap sampling and random subspace method to generate the initialized neural ensemble and maintains the randomization in the evolving stage, diversity in the ensemble is encouraged/kept in ERNE. Evolutionary algorithm with negative correlation learning is adopted to search for a population of diverse and accurate individual NNs that collectively solve a problem. In negative correlation learning, the individual networks are trained simultaneously, rather than independently or sequentially. Evolving the ensemble with negative correlation learning emphasizes not only the accuracy of individual NNs but also the cooperation among different individual NNs and thus improve the generalization. In ERNE, as each member in the ensemble is learned from bootstrap sample of the training examples, which typically omits $1/e \approx 37\%$ of the

training examples, out-of-bag (OOB) estimation, based on recording the votes of each member on those training examples omitted from its bootstrap sample and aggregating the votes for each training examples for an estimation of the generalization error, serves as another benefits of the algorithm.

The aim of this paper is to apply ERNE to the prediction of the Ames test mutagenicity of chemicals. The rest of this paper is organized as follows: The proposed algorithm is present in Section 2. Experimental results and discussion are reported in Section 3. Finally, Section 4 concludes the paper with future work.

2 Evolving Random Neural Ensembles with Negative Correlation Learning (ERNE)

It is widely believed that the success of ensemble algorithms depends on the accuracy and diversity among these base classifiers [12]. In general, the individual classifiers in ensemble are designed to be accurate and diverse among each other. For example, Bagging relies on bootstrap that produces different subsets of the training data; Ensemble of features employs different features instead of training data to generate diverse ensemble [8]. Random forests [13] combines Bootstrap sampling and random subspace method to generate more diverse ensembles. However, the predictions of random forests may fluctuate because of randomization of data and features simultaneously. Although the disadvantage of this could be slightly offset by including more and more decision trees in the ensemble, this of course leads to extended training times and more resources consumed.

The existing methods, random sampling of data and features, may promote the diversity but degrade the accuracy. How to improve the accuracy and simultaneously maintain the diversity for individual ensemble members to make sure that the obtained ensemble is both accurate and diverse is a key factor for ensemble algorithms. Chen et. al [1] proposed ERNE algorithm that offers a natural way to optimize accuracy and simultaneously maintain the diversity among the individuals in the ensemble. In this algorithm, randomization of both data and features has been adopted/kept to generate/maintain the diversity in the ensemble. Evolutionary ensemble with negative correlation learning provides the opportunities for these individual NNs to negatively correlated with each other and thus improve the accuracy of these individual NNs.

ERNE firstly generates an initial population of Neural Networks (NNs), each of which is trained on bootstrap of training data and random feature subspace. Then the diverse population is evolved with negative correlation learning to improve the accuracy of individual NN.

The whole process could be illustrated as following.

1. Sampling the original training set and obtain M replications of training set $\{B_i\}_{i=1}^M$.
2. Generate an initial population of M Neural Networks (NNs), the number of hidden nodes for each NN, $n_i (i = 1, \dots, M)$ is specified randomly restricted

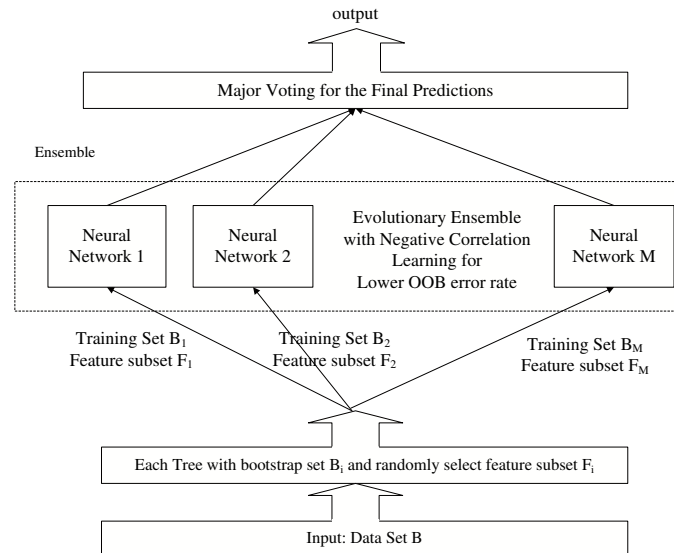


Fig. 1. The architecture of ERNE [1]

by the maximal number of hidden nodes. The random initial weights are distributed uniformly inside a small range.

3. Train each NN on each bootstrap set B_i with randomly selected feature subset $\{F_i\}_{i=1}^M$ for a certain number of epochs that is proportional to the number of hidden number of neural network using negative correlation learning and calculate the out-of-bag estimation as the ensemble fitness function.
4. In each generation, randomly choose s NNs to create offspring NNs¹. For each offspring s_i , evolve each s NNs with Gauss mutation², and train it with its corresponding parent's bootstrap set B_i and feature subset F_i . s is specified by the user.
5. Compare the fitness of each s_i NN with their respective parents and include the better one in the population and recalculate the out-of-bag error as the fitness.
6. Go to the next step if the maximum number of generations has been reached. Otherwise, and go to Step 3.
7. Combining the population to form the ensembles.

There are four advantages of this algorithm: (1) Ensemble of different data subset and feature subset promotes the diversity among individual classifier in the ensemble. (2) Evolving the individual NN in the ensemble helps to improve

¹ Each individual, selected to be mutated with equal probability, reflects the emphasis on evolving a diverse set of individuals.

² Add Gauss noise to the weight vector of neural network. The parameter of Gauss noise is: $mean = 0$ and $variance = \mu$, will be specified manually.

the accuracy of these NN. (3) Negative correlation learning enables these individual NNs in the ensemble correlated with each other and improves the generalization performance. (4) It generates an internal unbiased estimate of the generalization error, OOB, as the NN ensemble building progresses.

2.1 Negative Correlation Learning

Negative Correlation Learning (NCL), a successful neural network ensemble learning technique developed in the evolutionary computation literature [11,14], has shown a number of empirical successes and varied applications, including regression problems [15] and classification problems [16]. It has consistently shown promising results compared with other techniques like Mixtures of Experts, Bagging, and Boosting [11,17].

NCL introduces a correlation penalty term into the error function of each individual network in the ensemble so that all the networks can be trained simultaneously and interactively on the same training data set. Liu et al. [16] implemented NCL by gradient descent method for training neural network. In fact, negative correlation learning provides a novel way to decompose the learning task of the ensemble into a number of subtasks for different individual networks.

2.2 Out-of-Bag Fitness Evaluation

In ERNE, out-of-bag (OOB) estimation error is taken as the objective to be optimized. As each member in the ensemble is learned from bootstrap sample of the training examples, which typically omits $1/e \approx 37\%$ of the training examples. The out-of-bag estimate is based on recording the votes of each member on those training examples omitted from its bootstrap sample and aggregating the votes for each training examples for an estimation of the generalization error. Out-of-bag estimates is proposed as an ingredient in estimates of generalization error, which has been empirically supported by [18] that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. However, out-of-bag estimate requires considerably less time than the 10-fold cross-validation.

3 Experiments

In this section, we shall evaluate ERNE for the Ames test problem by comparing with some traditional classifiers, i.e. classification and regression tree (CART), and multilayer perceptions (MLP), as well as their corresponding ensemble learning algorithms respectively, i.e. Bagging of CART/MLP, Adaboost.M1 of CART/MLP and random forests/MLP.

3.1 Chemical Data Set

The chemical data used in this study were obtained from a paper of Kirkland et al. [19]. The total number of chemicals is 691, which consists of 357 chemicals with positive results from Ames tests and 334 chemicals with negative results.

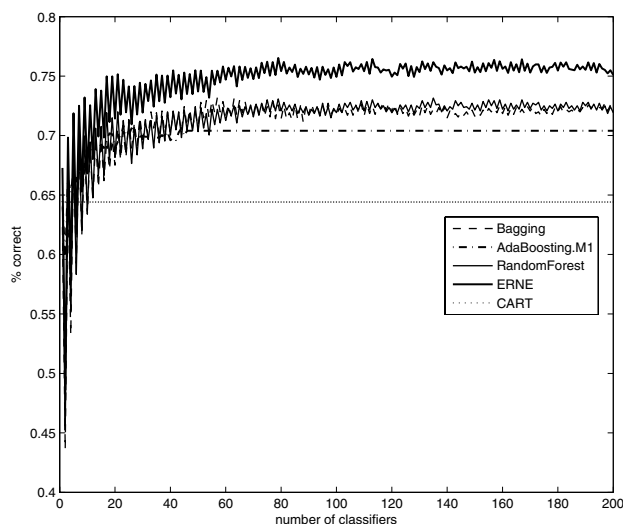


Fig. 2. Performance comparison of ERNE against CART and three other tree-based ensemble approaches in predicting the Ames test mutagenicity based on two-fold cross validation. The differences between ERNE and other classifiers are significant at the 5% significance level, see Table 2.

The variables that are used as features to build models are a set of 197 descriptors, which consist of atom/fragment counts, graph descriptors, topological descriptors and chemical structural descriptors [7].

3.2 Experimental Setup

As we know, the experimental results depend on the partitions of data set. In this paper, two-fold cross validation, allowing a sufficient test set to estimate the generalization error, is employed to evaluate these algorithms.

We preprocess all 197 features by normalizing them into values between 0.0 and 1.0. The network we used in this paper is three layer feedback NN. The number of hidden nodes will be initialized randomly but restricted in the range 3 to 8. Initial connection weights for individual NNs in an ensemble are randomly chosen. The parameter λ is set to 0.8 and the variance of Gaussian mutation is 0.1. The parameters in use are set to: the population size M (varied from 1 to 200), the number of offspring s ($\max[20, M]$), the number of generations (100). These parameters are chosen after some preliminary experiments. They are not meant to be optimal.

3.3 Experimental Results

The MLPs used in Bagging, Adaboosting.M1 and random MLPs are three-layer feedback NN with five hidden nodes. These MLPs are trained using scaled conjugate gradient (SCG) algorithm for 200 epoches. Figure 2 and 3 show the results

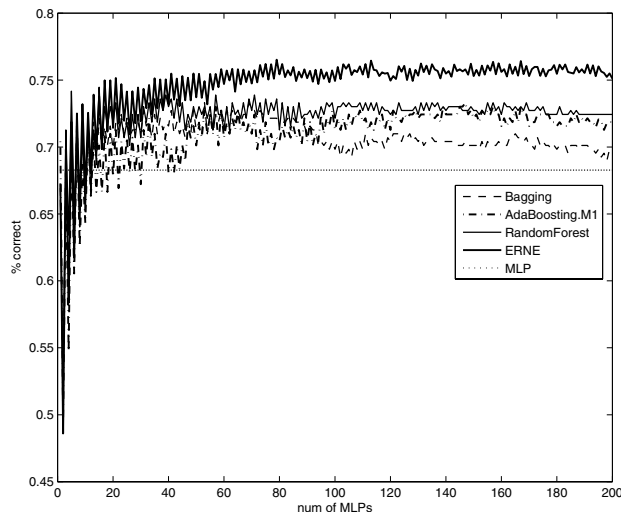


Fig. 3. Performance comparison of ERNE against MLP and three other MLP-based ensemble approaches in predicting the Ames test mutagenicity based on two-fold cross validation. The differences between ERNE and other classifiers are significant at the 5% significance level, see Table 2.

Table 1. Comparison Among ERNE with classification and regression tree (CART) and MLP based ensembles in terms of average cross validation error for Ames data set. The results are averaged on 2-fold cross-validation, respectively.

	% error ERNR RMLPs/RF		Bagging	Adaboost	MLP/Tree
MLP	24.90	27.57	30.46	28.14	31.74
CART	-	28.08	27.57	29.60	35.59

of ERNE over 40 independent cross validations. In each figure, we record the performance of these algorithms with respect to the size of the ensemble, i.e. the number of classifiers in this ensemble. For comparison, Table 1 lists the the average errors of all kinds of classifiers over the 40 runs and Table 2 gives the result of two-tail pared t-test in terms of prediction error between ERNE and other classifiers.

From Figure 2 and 3, ERNE consistently outperforms other algorithms in terms of cross validation error. This is understandable since the performance of random forests is better than or similar as the other ensemble algorithms in most of the cases, and ERNE maintains good diversity by adopting bootstrap and random feature subspace, which is similar as random forests, and evolves the ensemble to optimize the accuracy and cooperation. Generally speaking, ERNE will perform no worse than random forests.

In the preliminary analysis for Ames test problem, random forest approach performs best among a variety of approaches [7], including Decision Tree (C4.5), Naive Bayesain, K-NN, Logistic Regression, support vector machine (SVM),

Table 2. Results (P value) of two-tail pared t-test in terms of prediction error between ERNE and one other classifier

Methods vs ERNE	RMLPs/RF	Bagging	Adaboost	MLP/Tree
MLP	0.0376	0.0168	0.0093	0.0001
CART	0.0336	0.0236	0.0084	0.0000

Bagging and Boosting. The state-of-the-art performance of random forest approach is largely due to that it uniquely adopts both bootstrap of data and feature subset selection, which is particularly of value for handling a data set with a large amount of feature variables [8]. ERNE not only keeps the benefits of random forests but also improves its performance by optimizing the accuracy. The superiority of ERNE over random forests can be observed in the experimental results. In our experiments, we found that Adaboost.M1 of trees sometimes will overfit when adding more and more trees in the ensemble, which can be found in Figure 2 but for neural network ensemble, Adaboost.M1 algorithm behaves better when adding more and more MLPs in the ensemble. Another interesting point to say is that MLP-based ensemble, (Bagging, Adaboosting and random MLP), is not so stable as tree based ensemble. This is because tree can be seen as a deterministic algorithm and MLP would output differently given the same input because of randomization of initialization of weight vectors.

In this experimental results, all of the NNs are ensembled to constitute the final classifier. However, from Figure 2 and 3, ensembling some effective combination of NNs would be better than ensembling all of them. This will be considered as the future work.

The following two reasons might explain why the performance of our algorithm is better than that of others.

- ERNE generates a diverse ensemble. Firstly, bootstrap sampling of data and random feature subsets generate a diverse ensemble in the initial population, which inherits the merits of random forests. Secondly, in the evolving stage, the diversity is maintained by only mutating the weight of individual NN but not changing the bootstrap of data and the feature subset used by this individual NN.
- Evolving ensembles with negative correlation learning has a potential to simultaneously optimize both the accuracy and cooperation of the existing individual NNs in the ensemble, resulting in reducing the generalization error. The potential is indicated by our empirical results of this study.

4 Conclusions

This paper describes a novel machine learning ensemble approach, called ERNE, to building an *in silico* model for the prediction of the Ames test mutagenicity. *In silico* models serve major roles in reducing, refining and replacing animal testing for the risk assessment of the safety of chemicals. ERNE was developed based

on neural networks and evolutionary algorithms. Technically, it firstly combines the method of bootstrap sampling on training data with the method of random subspace feature selection to ensure diversity in creating individuals within an initial ensemble. Secondly, while evolving individuals within the ensemble, it makes use of the negative correlation learning, enabling individual NNs to be trained as accurate as possible while still manage to maintaining them as diverse as possible. Finally, it takes out-of-bag estimation as the fitness functions of individual NNs, which potentially enhances the generalization capabilities of individual NNs. Consequently, the resulting individuals in the final ensemble are capable of cooperating collectively well to achieve better generalization of prediction.

Empirical experiments have been carried out in this paper to evaluate ERNE on the Ames test mutagenicity prediction problem in comparison with other ensemble algorithms. ERNE has shown promising performance. The reasons to explain the superiority of ERNE are also given in this paper.

An immediate future work regarding the algorithm is to develop an ensemble subset selection method to choose a subset of individual NNs rather than all of individuals in the population to form the final ensemble, The hope is to improve the generalization performance and reduce computational complexity. We would also like to carry out some work to further understand the trade-off between accuracy and diversity of individual NNs in an ensemble in the context of improving overall generalization performance of ERNE.

For Ames test mutagenicity and other similar applications in chemoinformatics, the common characteristic of these problems is that there are plenty of features, each of which denotes a descriptor (graph descriptor, topological descriptor, chemical structural descriptor and so on). Although ensemble approaches, including ERNE, could achieve a high predict accuracy, they lack interpretability and thus make them less interesting from the viewpoint of toxicity decision making. The following work will focus on interpretability of approaches, e.g. feature selection and decision rules extraction.

Acknowledgment

This work is partially supported by a Dorothy Hodgkin Postgraduate Scholarship to the first author.

References

1. Chen, H., Yao, X.: Evolutionary random neural ensemble based on negative correlation learning. In: Proceedings of IEEE Congress on Evolutionary Computation (CEC'07). (2007) submitted.
2. He, L.N., Jurs, P.C., Custer, L.L., Durham, S.K., Pearl, G.M.: Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chemical Research in Toxicology* **16** (2003) 1567–1580

3. Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., Xie, Q.A., Tong, W.: Three new consensus qsar models for the prediction of ames genotoxicity. *Mutagenesis* **19** (2004) 365–377
4. Mahe, P., Ueda, N., Akutsu, T., Perret, J.L., Vert, J.P.: Graph kernels for molecular structure-activity relationship analysis with support vector machines. *Journal of Chemical Information and Modeling* **45** (2005) 939–951
5. Zhang, Q.Y., de Sousa, J.A.: Random forest prediction of mutagenicity from empirical physicochemical descriptors. *Journal of Chemical Information and Modeling* **47**(1) (2007) 1–8
6. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* **40**(2) (2003) 139–157
7. Li, J., Dierkes, P., Gutsell, S., Stott, I.: Assessing different classifiers for in silico prediction of ames test mutagenicity. In: A poster in the 4th Joint Sheffield Conference on Chemoinformatics. (2007) submitted.
8. Ho, T.K.: The random subspace method for constructing decision forests. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **20**(8) (1998) 832–844
9. Hansen, L.K., Salamon, P.: Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(10) (1990) 993–1001
10. Hashem, S.: Optimal linear combinations of neural networks. *Neural Networks* **10**(4) (1997) 599–614
11. Liu, Y., Yao, X.: Ensemble learning via negative correlation. *Neural Networks* **12**(10) (1999) 1399–1404
12. Brown, G., Wyatt, J., Tino, P.: Managing diversity in regression ensembles. *Journal of Machine Learning Research* **6** (2005) 1621–1650
13. Breiman, L.: Random forests. *Machine Learning* **45**(1) (2001) 5–32
14. Liu, Y., Yao, X.: Simultaneous training of negatively correlated neural networks in an ensemble. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **29**(6) (1999) 716–725
15. Yao, X., Fischer, M., Brown, G.: Neural network ensembles and their application to traffic flow prediction in telecommunications networks. In: *Proceedings of International Joint Conference on Neural Networks*. (2001) 693–698
16. Liu, Y., Yao, X., Higuchi, T.: Evolutionary ensembles with negative correlation learning. *IEEE Transaction on Evolutionary Computation* **4**(4) (2000) 380–387
17. McKay, R., Abbass, H.: Analyzing anticorrelation in ensemble learning. In: *Proceedings of 2001 Conference on Australian Artificial Neural Networks and Expert Systems*. (2001) 22–27
18. Breiman, L.: Out-of-bag estimation. Technical report, Stanford University (1996)
19. Kirkland, D., Aardema, M., Henderson, L., Muller, L.: Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens i. sensitivity, specificity and relative predictivity. *Mutation Research* **584** (2005) 1–256