

# MULTI-VERSION SYSTEMS OF NEURAL NETWORKS FOR PREDICTING THE RISK OF OSTEOPOROSIS

Wenjia Wang, Derek Partridge and Sarah Rae\*

Dept. of computer science, University of Exeter, Exeter, EX4 4PT, UK

w.wang@ex.ac.uk, Derek@dcs.ex.ac.uk

\*Osteoporosis research center, Royal Devon and Exeter Hospital, UK.

## Abstract

This paper describes the methodology of developing multi-version systems using neural networks in the hope of improving their performance and reliability. However, a system implemented by simply combining  $N$  neural nets may not necessarily deliver a better result than the individual versions alone. A critical factor to success is the diversity among these versions, which is high probability that the system will avoid coincident failures and therefore exhibit increased reliability. Coincident-failure diversity (CFD) is described as a specific measure of the diversity quantitatively. The approach of Multi-Net System(MNS) has been applied to predict the risk of osteoporosis for female patients. The performance of the MNS showed with ROC curves are considerably better than that of the individual nets in the system and also Logistic regression.

**Keywords:** Multi-version systems, neural networks, diversity, ROC curve, osteoporosis

## 1. INTRODUCTION

The technology of neural networks has been successfully applied to solve various problems which prove difficult to solve using conventional methods. However, the reliability of neural net systems has become one of the most important issues associated with the technique ever since it emerged. It is generally accepted that a single trained neural network or one selected from a number of nets specially trained for a given problem, may not perform reliably due to the nature of neural net learning and subsequent execution. Therefore, improving this weakness has being an essential objective of neural computing research.

The concept of Multi-Version System (MVS)s has emerged, but seldom been implemented in traditional software programming. It can be used in neural computing to enhance the performance and reliability of neural networks.

This paper will describe a methodology for developing multiversion systems with Multi-Layer Perceptron (MLP) neural networks. In addition, some measurements are defined to evaluate a system's overall performance in terms of generalisation probability, diversity and reliability. The approach of the MVS is applied to predict the risk of osteoporosis with the real data. Comparisons are presented of the overall performance between the MVS approach, single nets, and Logistic Regression.

## 2. MULTI-VERSION SYSTEMS

A general *Multiversion System* (MVS) is commonly described as the one in which the basic functionality of the system for a given problem is redeveloped in a variety of multiple individual programs (or versions). A decision strategy then is needed to determine the overall outcome of the system from the performances of its individual versions.

The idea of multiversion software engineering had been suggested as an approach to reliability enhancement, studies both analytical and empirical were conducted to examine and test what appeared to be a plausible hypothesis: if  $N$  versions of a system are constructed independently, they will not make the same errors, and the faults occurring in the different versions will be unrelated. And thus a majority decision over the  $N$  computations will be more reliable than any one of the  $N$  versions alone. Initial studies focused on defining "independence" and determining if independence of failure actually occurred[1]. The results rejected the hypothesis: in other words, independently developed versions will fail dependently, which led to pessimistic prognostications for the future of  $N$ -version programming [2].

As part of a comprehensive conceptual model of coincident failures in multiversion programming Littlewood and Miller[3] questioned the pivotal significance of independent failure behaviour, and proposed that the essential matter is *diversity*. In order to address diversities they defined measures of the system,

such as  $E(\Theta)$  –the probability that a randomly selected version  $\Theta$  from  $N$  versions fails on a randomly selected input,  $E(\Theta^2)$  – the probability that two versions selected at random both fail on a random input, etc.

Partridge et al. [4,5,6] extended Littlewood and Miller's work and proposed a methodology of implementing an MVS by using neural networks, and then developed the corresponding statistical models.

### 3. MULTI-VERSION SYSTEMS of NEURAL NETS

#### • Multi-Neuralnet System(MNS)s

In contrast to the conventional software engineering, developing a version of neural network does not require further programming once a simulation programme (or package) is developed. Therefore the task of training a variety of neural networks is not only easier and faster (even though the massive parallel operation of neural networks is simulated on a sequential computer), but also cheaper. It can be done by altering their initial conditions, learning algorithm parameters and training data. Consequently a Multi-version system of neural networks, - Multi-Neuralnet System (MNS), can be easily implemented.

Many researchers, such as Hansen et al.[7], Krogh et al.[8] and Ghosh et al.[9], Opitz et al.[10], etc. have successfully applied the idea by combining a number of trained neural nets, named as ensemble, to improve the accuracy. But some other experienced contrary applications. An example can be seen in [11] where the committees of the trained nets were constructed and the results they obtained were no better than those of the corresponding individual nets in the committee. [8] tried to explain the reason behind these successes and failures by the concept of ambiguity which was later defined as diversity by Opitz et al.[10]. However, their measures of the diversity appeared still lack of reflecting the probability of coincident failures which is, we believe, more important than the other diversity. In fact, diversity has long been recognized as a critical factor to the success of multi-version systems in conventional software engineering and therefore some developed measures in that field can be adopted for our purpose.

This research uses CFD (coincident-failure diversity) defined by Partridge et al.[4,5,6] as one of diversity measures to guide the construction of multi-version systems of neural networks.

#### • Construction of MNSs

Multi-Layer Perceptron (MLP)s trained by the backpropagation learning algorithm have been used to create multi-net systems. The flexibility in training MLPs provides options on strategies, such as altering initial and terminal conditions, parameters of learning algorithms, changing structure of the net and manipulating training set etc. for producing different versions of the nets. The nets trained with specific set-ups are placed into their corresponding category of net family, then a multi-net

system can be constructed by the nets selected from a single net family or all families with pre-defined criteria, or without selection - simply taking one complete net-family as an MNS.

### 4. PERFORMANCE ASSESSMENT

#### • A Decision-Making Strategy

For a multiversion system composed of  $N$  nets a decision strategy must be applied in order to yield a final outcome of the system. Quite a few [6, 9] can be used. For dichotomous categorisation problems, the strategy of majority voting appears to be a reasonable choice. The operation is performed by counting the decisions of the individual versions in the system, and then making the final decision depending on the category which possesses the majority votes. It can be formulated by equation (1).

$$v_i = \begin{cases} 1 & \text{if } \sum_{j=1}^N z_{ji} \geq \left\lceil \frac{N}{2} \right\rceil \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where  $z_j$  denotes the Boolean output, {0,1}, of version  $j$  after squashing its raw output with a threshold function for input pattern  $i$ ,  $v_i$  the decision of the  $N$ -version system, and  $N$  should be normally set up as an odd number in this case.  $\lceil N/2 \rceil$  represents the ceiling operation which returns the nearest integer larger than  $N/2$ , i.e. the required smallest simple majority number.

#### • Diversity measures and performance assessment

Assume that  $M$  patterns in a test set are presented to each of  $N$  versions in a multiversion system. The outputs from them are classified either success or failure. Let  $m_q$  denote the number of test patterns that fail on  $q$  versions and then the probability,  $p_q$ , that  $q$  versions within the system will fail simultaneously on a randomly chosen input from a population of test patterns is defined as  $p_q = m_q/M$ . The probability,  $p_0$ , is that all test patterns fail on no version (or success on all versions, in other words), thus CFD is then calculated with equation (2) (the details of derivation in [4]) and  $CFD \in [0,1]$ .  $CFD=0$  indicates either all failures are same in all versions—hence no diversity, or no test failure at all versions (an ideal case), i.e. all versions are perfect and identical—hence no diversity. In this latter case there is also no need for diversity since a single perfect version is enough. Unfortunately, no one has found a way to produce such a perfect version for most real problems yet.  $CFD=1$  when all test failures are unique to one version, i.e. maximum diversity.

$$CFD = \begin{cases} \frac{1}{1-p_0} \sum_{q=1}^N \frac{N-q}{N-1} p_q, & \text{if } p_0 < 1 \\ 0, & \text{if } p_0 = 1 \end{cases} \quad (2)$$

The performance of the systems or individual nets are assessed with the probability of generalisation,  $p(G)$ , which is defined as a ratio of the number of success

patterns over the number of total test patterns. The reliability of the systems is quantitatively evaluated by a number of measures defined by [3] including  $E(\Theta)$ ,  $E(\Theta^2)$ ,  $p(1/3)$  -the probability that at least 1 out of 3 randomly selected versions from the system is correct and  $p(1/N)$ , 1 out of  $N$  versions correct.

In addition, ROC (Receiver Operating Characteristic) curve, a commonly used means for assessing the performance of a conventional classifier or predictor for dichotomous problems, is also adopted for evaluating MNSs. A ROC curve is plotted by the probability of sensitivity,  $p(sen)=(\text{number of the correctly predicted patterns for the cases})/(\text{the total number of the cases})$ , against  $1-p(spe)$ ,  $p(spe)=(\text{number of the correctly predicted patterns for the non-cases})/(\text{the total number of the non-cases})$ , as varying the value of the decision threshold. Based on this concept an alternative measure of reliability of the system performance is defined as the Euclidean distance,  $D(sen,spe)$ , from the points along the curve to the upper left corner (an optimal solution point) on the ROC plan. Thus the point with the shortest distance is then considered as the best overall performance point of the system.

$$D(sen,spe)=\sqrt{(1-p(spe))^2+(1-p(sen))^2} \quad (3)$$

## 5. OSTEOPOROSIS PROBLEM

The methodology of developing Multi-Neuralnet Systems has been initially applied to a number of simple problems, such as Launch Interceptor Conditions problems, as illustrations[4,5,6], and achieved the expected satisfactory results. Then a research project on a medical problem, aimed to predict the risk of osteoporosis among the patients presented in clinics, has being conducted collaborating with the experts of osteoporosis in hospital to further investigate the capability of MNS for dealing with real problems.

Osteoporosis is a systemic skeletal disease characterised by low bone mass and microarchitectural deterioration of bone tissue, with a consequent increase in bone fragility and susceptibility to fractures. People are often unaware of developing osteoporosis since the early symptoms of the disease are not obvious and may not cause pain until osteoporosis disease is already progressed to the advanced stage and bones start to fracture. Osteoporosis affects millions of people worldwide every year. Women have about 4 times greater risk than men. Particularly affected are elderly women after menopause, who lose bone mass at a high rate because of the reduced oestrogen levels. Currently, there is no cure for osteoporosis and the best treatment is preventive measures, which requires early identification of the level of osteoporosis. The earlier the diagnosis, the more effective are preventive treatments. A variety of bone density(directly and indirectly) measuring devices (DEXA, Ultrasound Scanner, etc.[12,13]) could

serve for this purpose but can only be used efficiently on candidates who are considered at higher risk by GPs or consultants due to either availability of the resources, or financial constraints. It is inevitable in this situation that some patients at higher risk are not referred for screening while many patients are referred unnecessarily.

There is a real need, therefore, to develop a computer-aided tool which could be used to predict the risk of developing osteoporosis for potential patients and to help GPs or relevant consultants to make optimal decision. This is the aim of the research and the success of it will certainly benefit patients and improve the efficiency of the screening devices.

The data on 274 female patients were collected with the identical questionnaire forms listing more than 40 items involving the patient's personal facts, life style and relevant history, disease diagnosis and treatments etc. The diagnosis of the patients were confirmed by the results of the Ultrasound scanner [12] in terms of T-score. 20 factors, such as: age, weight, height loss, kyphosis, menopause, fractures, life style, such as alcohol consuming, smoking, exercise, and inactivity, and the relevant diseases and medications etc. are selected by the experts in the field as the inputs. The patients are classified in this research by their T-score of the scanning test into two categories, normal and abnormal (osteoporotic).

## 6. EXPERIMENT DESIGN

### • Mapping the problem onto the MLP net

A number of schemes were tested to find a suitable one for mapping the problem onto MLP net. 20 factors are designed as the input variables of an MLP net. The category of the patients: normal and abnormal encoded as 0 and 1, is taken as an output. Thus a MLP must have a structure of 20-H-1, H is the number of hidden neurons in single hidden layer, which was set up empirically to 5, 10 and 15 in the experiments.

### • Construction of the training sets

With less than 300 of data patterns available, the way of utilising of the data is critical for the net training and test. A mechanism described in the author's previous paper [6] is used for constructing a representative set and  $N$  subtraining sets. About 65% ( $n=180$ ) of the total data are taken to form a representative training set,  $Q$ . Then  $N$  sub-training sets are constructed with intersection rate  $\zeta$  varying from 1 to 0 as the size of subset,  $k$ , increasing from 20 to  $n$  with irregular steps. The remaining patterns are used for validation. The purpose is to force the individual nets learn different features of the problems and to create high diversity among the various versions.

### • Construction of Multi-Net Systems

The primary multi-version system is designed containing  $N$  ( $=9$ ) MLP neural nets which are trained with their corresponding sub-training sets at each specific  $k$ . Two

variations are selected for the training, i.e. initial weight conditions (W1 and W2) and terminal criteria (F1 - family one of nets which are trained to all or near all patterns correctly learned, and F2—family two whose nets are just trained to reach majority right). Without any selection all trained nets in their own categories are directly taken to form four types of multi-version systems can be obtained built for a given  $k$ , i.e. W1\_F1, W1\_F2, W2\_F1 and W2\_F2. Varying hidden units,  $H(=5, 10, 15)$ , will create some other sets of net families, e.g. Hx\_W1\_F1. Then a mixed MNS is composed of the nets selected from all the trained net families with a predefined criterion, a combined  $p(G)$  and the diversity.

#### • Test and evaluation of the systems

The constructed MNSs are then tested with a test set, the remaining data patterns (94, about 35%) other than the patterns in the representative training set. The systems are then evaluated by the measures mentioned in the earlier section.

### 7. THE COMPUTATIONAL RESULTS

The experiments were simulated with a software developed in the author's research group and some relevant items of the evaluation report for the MNS, H5\_W1\_F2, when  $k = 100$  ( $\zeta=0.2$ ), are listed in Table 1.

**Table 1 The measures for a multi-net system**

MNS: H5\_W1\_F2,  $N=9$ ,  $k=100$

$p(\text{BestNet})=0.7660$ ,  $p(\text{WorstNet})=0.6809$ ,  $\text{Mean}(p)=0.7198$ ,

$E(\Theta)=0.2801$ ,  $E(\Theta^2)=0.0784$ ,  $\text{Var}(\Theta)=0.1016$ ,

CFD = 0.5741,  $p(\text{Maj})=0.7766$ ,

$P(1/2)=0.8404$ ,  $P(1/3)=0.7765$ ,  $P(2/3)=0.7514$ ,  $P(1/9)=0.9255$

$\text{Mean}(p)$  denotes that the mean value of generalisation probability over all individual nets in the system.  $p(\text{BestNet})$  and  $p(\text{WorstNet})$  are the generalisation probability of the best net and the worst net in the system respectively.  $p(\text{Maj})$ , the probability of the MNS with majority-voting strategy. It is clear that the MNS with majority-voting strategy produced about 6% higher generalisation rate than the mean of the individual nets' performance. With a relatively higher diversity (CFD=0.5741) this MNS is expected to have a higher reliability to cope with the complex cases compared with that single net and the MNSs with lower diversity. We have observed that the MNSs with lower CFD, their  $p(\text{Maj})$  are worse than the best nets in their own systems, which demonstrates a hypothesis that simply adding the number of the versions to the system may not necessarily lead to improvement of the overall performance and sometimes even worse. This also indicates the importance of the diversity, and the necessity of applying the selecting rules when constructing multi-version systems.

All the nets trained under various conditions are put into a pool as candidates for selection, three nets with higher

combined value of  $p(G)$  and coincident-failure diversity are chosen to construct an MNS. Then this MNS is tested again with the same test set as before. The results are listed in Table 2.

**Table 2. Evaluations of the MNS of the selected nets**

| Version   | $p(\text{spe})$ | $p(\text{sen})$ | $p(G)$ | No.Vns | Co.Fl | $p(\text{fail})$ |
|---|-----------------|-----------------|--------|--------|-------|------------------|
| 1   | 0.476           | 0.904           | 0.809  | 0/3    | 58    | 0.6170           |
| 2   | 0.571           | 0.877           | 0.808  | 1/3    | 20    | 0.2128           |
| 3   | 0.143           | 0.857           | 0.776  | 2/3    | 5     | 0.0532           |
| mean  | 0.397           | 0.879           | 0.798  | 0/3    | 11    | 0.1170           |
| $p(\text{BestNet})=0.809$ , $p(\text{WorstNet})=0.776$ , $\text{Mean}(p)=0.798$<br>$E(\Theta)=0.1801$ , $E(\Theta^2)=0.0353$ , $\text{Var}(\Theta)=0.1171$ ,<br>CFD = 0.6251, $p(\text{Maj})=0.8298$<br>$P(1/2 \text{ correct})=0.8652$ , $P(1/3)=0.8829$ , $P(2/3)=0.8298$ |                 |                 |        |        |       |                  |

Where, "No.Vns" denotes that the number of Versions(out of  $N=3$ ) fail(s) the test and "Co.Fl", the number of the patterns which fails on that number of versions, e.g. 1/3, 20 (in the third row) indicate that 20 test patterns are failed on exactly 1 out of 3 versions(1/3). It is obvious that  $p(\text{Maj})=0.8298$  is higher than that of any individual nets alone.

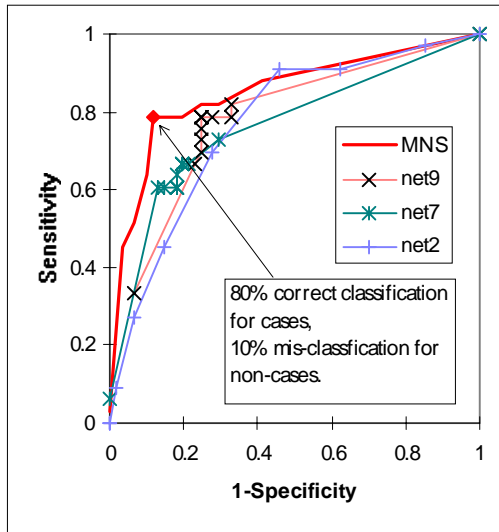
### 8. COMPARISON

The results obtained by the MNS approach has been compared with that of the Logistic Regression, one of the most commonly used conventional methods for classification or prediction problems in medical field. Table 3 (a) is the regression results with 180 patterns,  $p=0.0011$ ,  $df=20$ , and (b) is the test results with the same test set as that used by the MNS.

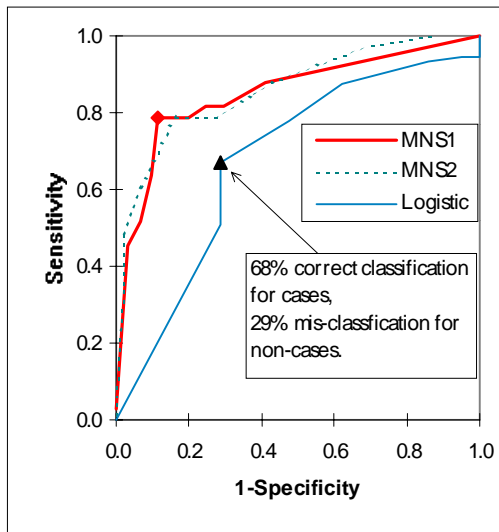
**Table 3. The results of the logistic regression**

| (a) Regression      |           |     |             | (b) Test    |    |         |
|---------------------|-----------|-----|-------------|-------------|----|---------|
|                     | Predicted |     | Correct     | Predicted   |    | Correct |
| Observed            | 0         | 1   | (%)         | 0           | 1  | (%)     |
| 0                   | 20        | 34  | 37          | 3           | 19 | 14.3    |
| 1                   | 15        | 111 | 88.1        | 8           | 64 | 87.7    |
| Overall performance |           |     | <b>72.8</b> |             |    |         |
|                     |           |     |             | <b>71.3</b> |    |         |

The comparison was also done by plotting the ROC curves of these two approaches. Figure 1 compares the performance of the MNS and the three best individual nets in the system, which clearly indicates that the MNS has much higher correct prediction rate,  $p(\text{sen})$ , for the osteoporotic cases than that any of the individual nets at the same mis-prediction rate,  $p(1-\text{spe})$ , for the non-cases. Figure 2 shows the comparison between the MNSs and the Logistic Regression. The MNSs significantly outperformed the Logistic regression approach in all conditions of decision threshold. The best performance points, according to equation (3), are marked by the solid spots on their corresponding curves, i.e.  $p(\text{spe})=0.80$ ,  $p(\text{sen})=0.90$  and  $D(\text{sen}, \text{spe})=0.24$  for the MNS1 (MNS2 is similar) and  $p(\text{spe})=0.68$ ,  $p(\text{sen})=0.71$  and  $D(\text{sen}, \text{spe})=0.46$  for the Logistic regression.



**Figure 1.** ROC curves of the MNS and three represent individual nets in the system.



**Figure 2.** ROC curves of two MNSs and the Logistic regression

## 9. CONCLUSION

This paper has described the research of adopting the concept of Multi-Version Systems (MVS) in conventional software engineering to construct multi-net systems in order to (i) enhance the generalisation ability, (ii) improve the reliability of neural computing. However, the more important concern we found is that not every MNS can guarantee a better performance. It is the diversity (plus relatively high generalisation) that does. The experiments we conducted on osteoporosis problem have demonstrated that an MNS with higher diversity will have higher reliability than the MNSs with lower diversity. The results achieved by the MNSs with high CFD are much better than the average of performance of the individual trained nets and also significantly better than the Logistic regression.

However, it should be pointed out that the experiments were conducted with the quite small quantity of the data and the results of this stage needs to be further improved before for clinical practice. Moreover, how to achieve a maximum diversity is still not a clearly understood issue, which certainly deserves further investigation. These initial results are, nevertheless, encouraging and indicate that the approach of multi-version systems of neural networks is effective for improving the overall performance of neural computing.

Acknowledgement: This research is supported by the EPSRC (UK) grant, FDR0802.

## References

- [1] Eckhardt, D. et al. A theoretical basis for the analysis of multiversion software subject to coincident errors. IEEE Trans. Software Eng. vol. SE-11, no. 12, pp. 1511-1517, 1985.
- [2] Brilliant, S. et al: Analysis of faults in an N-version software experiment. IEEE Trans. Software Eng. vol. 16, no. 2, pp. 238-247, 1990.
- [3] Littlewood, B. & Miller, D. Conceptual modelling of coincident failures in multiversion software. IEEE Trans. Software Eng. vol. 15, no. 12, pp. 1596-1614, 1989.
- [4] Partridge, D. & Krzanowski, W. Software diversity: practical statistics for its measurement and exploitation. Information and Software Technology (in press).
- [5] Partridge, D. et al: Engineering multiversion neural-net systems. Neural Computation, vol. 8, no. 4, pp. 869-893, 1996
- [6] Wang, W. & Partridge, D. Multi-version systems and reliability of neural computing. CACUK'97, pp141-147, Manchester, UK. Sep. 1997.
- [7] Hansen, L. et al. (1990) Neural network ensembles. IEEE Trans. Patterns Analysis and Machine Intelligence, Vol. 12(10), pp993-1001.C. 1990.
- [8] Krogh, A. et al. Neural network ensembles, cross validation, and active learning. In Touretzky, D. editors, Advances in neural information processing systems. Vol. 7, MIT Press. 1995.
- [9] Ghosh, J. et al. Integration of neural classifiers for passive sonar signals. In Leondes, C. editor, Control and Dynamic systems-Advances in theory and Applications, Vol. 77, pp301-338, Academic Press. 1996.
- [10] Opitz, D. et al. Generating accurate and diverse members of a neural network ensemble. In Touretzky, D. et al editors, Advances in neural information processing systems. Vol. 8, MIT Press. 1996.
- [11] Lippmann, R. et al. Predicting the risk of complications in coronary artery bypass operations using neural networks. Neural Information Processing Systems, Vol.7 pp1055-62, 1994.
- [12] Langton, C. The Role of ultrasound in the assessment of osteoporosis. Clinical Rheumatology, vol. 13, pp13-17, 1994.
- [13] Schott, A. et al. Ultrasound in discriminates patients with hip fracture equally well as Dual Energy X-Ray Absorptiometry and independently of bone mineral density. J. of Bone and Mineral Research. vol. 10, no. 2, pp243-249, 1995.