

Mineral Potential Mapping Using Bayesian Learning for Multilayer Perceptrons

Andrew Skabar

Received: 14 November 2005 / Accepted: 19 March 2007 / Published online: 24 August 2007
© International Association for Mathematical Geology 2007

Abstract Multilayer perceptrons (MLPs) can be used to discover a function which can be used to map from a set of input variables onto a value representing the conditional probability of mineralization. The standard approach to training MLPs is gradient descent, in which the error between the network output and the target output is reduced in each iteration of the training algorithm. In order to prevent overfitting, a split-sample validation procedure is used, in which the data is partitioned into two sets: a training set, which is used for weight optimization, and a validation set, which is used to optimize various parameters that can be used to prevent overfitting. One of the problems with this approach is that the resulting maps can display significant variability which stems from (i) the (randomly initialized) starting weights and (ii) the particular training/validation set partition (also determined randomly). This problem is especially pertinent on mineral potential mapping tasks, in which the number of deposit cells is a very small proportion of the total number of cells in the study area. In contrast to gradient descent methods, Bayesian learning techniques do not find a single weight vector; rather, they infer the posterior distribution of the weights given the data. Predictions are then made by integrating over this distribution. An important advantage of the Bayesian approach is that the optimization of parameters such as the weight decay regularization coefficient can be performed using training data alone, thus avoiding the noise introduced through split-sample validation. This paper reports results of applying Bayesian learning techniques to the production of maps representing gold mineralization potential over the Castlemaine region of Victoria, Australia. Maps produced using the Bayesian approach display significantly less variability than those produced using gradient descent training. They are also more reliable at predicting the presence of unknown deposits.

A. Skabar (✉)

Department of Computer Science and Computer Engineering, La Trobe University, Bundoora,
3086 VIC, Australia
e-mail: a.skabar@latrobe.edu.au

Keywords Mineral potential mapping · Neural networks · Bayesian learning · Markov chain Monte Carlo (MCMC) methods

Introduction

There has recently been considerable interest in applying multilayer perceptrons (MLPs) to the task of mineral potential mapping (Brown et al. 2000; Porwal et al. 2003; Harris and Pan 1999; Skabar 2005a). The appeal of MLPs is their property of being universal function approximators; that is, they are capable of representing highly non-linear relationships to an arbitrary degree of accuracy (Cybenko 1989; Hornick et al. 1989). In practice, this property is both a blessing and a curse. It is a blessing because we know that the functional form is capable of representing the target function; but it is a curse because we must go to great lengths to ensure that the MLP has not been overfitted to the training data.

Several standard techniques exist for preventing overfitting: early stopping, in which training is stopped before overfitting begins to occur; structural stabilization, which controls the expressive capability of the MLP by means of limiting the number of hidden layer units; and weight regularization, in which an additional term, which penalizes large weights, is incorporated into the error function. As there is no precise way to specify a priori the optimal number of hidden layer units, the value of the weight regularization coefficient, or the optimal number of training iterations, these parameters must usually be optimized offline using a split-sample validation procedure which monitors the error on a set of holdout (validation) examples not used in the weight optimization process.

While split-sample validation is a satisfactory procedure on many classification and regression tasks, its use on mineral potential mapping tasks is problematic for several reasons. First, mineralization is a rare event, and consequently, the number of mineralized pixels will be a very small proportion of the total number of pixels in the study area. This in itself imposes difficulty for weight optimization, and removing some of these examples for validation purposes only exacerbates this difficulty. Second, split-sample validation gives a noisy estimate of the error because it depends on the particular training/validation set partition used. Moreover, the degree of noise is inversely related to the number of validation examples, and thus it will be high on mineral potential mapping tasks due to the small number of known deposits. In practice, the difficulty in determining optimal values for these overfitting avoidance parameters is further compounded by the fact that interdependencies exist between them. For example, optimal regularization coefficient values for a complex model (i.e. one with a large number of hidden layer units) might be different than those for a simpler model. Also, early stopping (if used) may impact on the choice of weight regularization coefficients. Furthermore, given suitable values for these parameters, there can still be considerable variance in networks trained from different starting weights.

In the context of mineral potential mapping, these problems manifest themselves in significant variability between maps produced in separate runs of the same algorithm on the same data. This is discomforting, because a map should depend fundamentally on the data itself, and not on incidental factors such as starting weights or

the particular training/validation set partition. This problem has drawn little attention in the mineral potential mapping literature.

In recent years, several important advances have been made in the area of MLP training, one of which has been in the area of Bayesian learning (MacKay 1992; Neal 1996; Denison et al. 2002). Preliminary results of applying Bayesian learning techniques to mineral potential mapping have been described in Skabar (2005b). In contrast to gradient descent methods, Bayesian learning techniques do not seek a single ‘best’ weight vector; rather, they infer the posterior distribution of weights given the data. Predictions are then made by integrating over this distribution. A distinct advantage of the Bayesian approach is that the optimization of parameters such as the weight decay regularization coefficient can be performed using training data alone, thus avoiding the uncertainty introduced through split-sample validation. In the context of mineral potential mapping, this is expected to lead to maps which display less variability than maps produced using gradient descent methods. Furthermore, the integrative nature of the Bayesian approach is expected to make it more robust against noise in the training data, resulting in maps which make more reliable predictions on test data than maps produced using gradient descent methods.

Bayesian Learning for MLPs

An MLP is a mathematical function parameterized by a set of numerical weights $w_1, w_2, \dots, w_N$, which we shall denote collectively as a weight vector $\mathbf{w}$. The Bayesian approach involves inferring the posterior distribution of weights, $p(\mathbf{w} | D)$, given data D . It considers a probability distribution function over the weight space. The predicted output corresponding to some input vector $\mathbf{x}$ is then obtained by performing a weighted sum of the predictions over all possible weight vectors, where the weighting coefficient for a particular weight vector depends on the posterior weight distribution. Thus,

$$\hat{y}^n = \int f(\mathbf{x}^n, \mathbf{w}) p(\mathbf{w} | D) d\mathbf{w}, \quad (1)$$

where $f(\mathbf{x}^n, \mathbf{w})$ is the MLP output, and $\hat{y}^n$ is the predicted value.

Because $p(\mathbf{w} | D)$ is a probability density function, the integral in (1) can be expressed as the expected value of $f(\mathbf{x}^n, \mathbf{w})$ over this density

$$\int f(\mathbf{x}^n, \mathbf{w}) p(\mathbf{w} | D) d\mathbf{w} = E_{p(\mathbf{w}|D)}[f(\mathbf{x}^n, \mathbf{w})] \simeq \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}^n, \mathbf{w}). \quad (2)$$

Thus, the integral can be estimated by drawing N samples from the density $p(\mathbf{w} | D)$, and averaging the predictions due to these samples. This process is known as Monte Carlo integration. The density $p(\mathbf{w} | D)$ can be estimated using the fact that $p(\mathbf{w} | D) \propto p(D | \mathbf{w})p(\mathbf{w})$, where $p(D | \mathbf{w})$ and $p(\mathbf{w})$ are known respectively as the likelihood and prior. We now discuss the calculation of $p(D | \mathbf{w})$ and $p(\mathbf{w})$.

Likelihood Calculation on Mineral Potential Mapping Tasks

Let $\mathbf{x}^n = (x_1^n, x_2^n, \dots, x_m^n)$ be the attribute vector describing pixel n . Suppose that example $\mathbf{x}^n$ with target value t^n is drawn randomly from the training set, and that t^n has a value of 1 if n contains a known deposit, and 0 otherwise. We assume that $f(\mathbf{x}^n, \mathbf{w})$ represents the conditional probability that pixel n contains one or more of the known deposits, given the attribute vector $\mathbf{x}^n$ (Skabar 2005a). Therefore, the probability that t^n equals 1 is $f(\mathbf{x}^n, \mathbf{w})$, and the probability that t^n equals 0 is $1 - f(\mathbf{x}^n, \mathbf{w})$. The probability of observing the correct target value is

$$p(t^n | \mathbf{x}^n, \mathbf{w}) = f(\mathbf{x}^n, \mathbf{w})^{t^n} (1 - f(\mathbf{x}^n, \mathbf{w}))^{1-t^n}. \quad (3)$$

Assuming that the examples are independent and identically distributed (i.i.d.), the probability of observing the correct target value for all examples (i.e. the likelihood) is given by

$$\begin{aligned} p(D | \mathbf{w}) &= \prod_n f(\mathbf{x}^n, \mathbf{w})^{t^n} (1 - f(\mathbf{x}^n, \mathbf{w}))^{1-t^n} \\ &= \exp\left(-\left(-\sum_n \{t^n \ln f(\mathbf{x}^n, \mathbf{w}) + (1 - t^n) \ln(1 - f(\mathbf{x}^n, \mathbf{w}))\}\right)\right). \end{aligned} \quad (4)$$

Prior Weight Distribution

The prior weight distribution, $p(\mathbf{w})$, is the weight distribution before observing any data, and reflects the prior knowledge that we have about the complexity of the MLP. To reflect the fact that we want it to be a smooth function (in order to reduce the risk of overfitting), we assume $p(\mathbf{w})$ to be Gaussian with zero mean and inverse square variance α , that is

$$p(w) = \left(\frac{\alpha}{2\pi}\right)^{m/2} \exp\left(-\frac{\alpha}{2} \sum_{i=1}^m w_i^2\right), \quad (5)$$

where m is the number of weights in the MLP. Because α controls the value of other parameters (i.e. the weights), it is referred to as a hyperparameter (Neal 1996).

A problem is that we usually do not know what variance to assume for the prior distribution, and for this reason it is common to set a distribution of values. Because α must be positive, a suitable form for its distribution is the gamma distribution

$$p(\alpha) = \frac{(a/2\mu)^{a/2}}{\Gamma(a/2)} \alpha^{a/2-1} \exp(-\alpha a/2\mu), \quad (6)$$

where a and μ are respectively the shape and mean of the gamma distribution (Neal 1996).

Because the prior depends on α , (1) should be modified such that it includes the posterior distribution over parameters and hyperparameters. Thus

$$\hat{y}^n = \int f(\mathbf{x}^n, \mathbf{w}) p(\mathbf{w}, \alpha | D) d\mathbf{w} d\alpha, \quad (7)$$

where $p(\mathbf{w}, \alpha | D) \propto p(D | \mathbf{w})p(\mathbf{w} | \alpha)p(\alpha)$.

Monte Carlo Integration

Monte Carlo integration requires obtaining samples from the posterior distribution. The objective is to sample preferentially from the region where $p(\mathbf{w}, \alpha | D)$ is large. The Metropolis algorithm (Metropolis et al. 1953) achieves this by generating a sequence of vectors in such a way that each successive vector depends on the previous vector as well as having a random component, that is, $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + \varepsilon$, where ε is a small random vector. Preferential sampling is then achieved using the criterion

$$\begin{cases} \text{if } p(\mathbf{w}_{\text{new}} | D) > p(\mathbf{w}_{\text{old}} | D) & \text{accept,} \\ \text{if } p(\mathbf{w}_{\text{new}} | D) < p(\mathbf{w}_{\text{old}} | D) & \text{accept with probability } \frac{p(\mathbf{w}_{\text{new}} | D)}{p(\mathbf{w}_{\text{old}} | D)}. \end{cases} \quad (8)$$

The difficulty in using the Metropolis algorithm to estimate the integrals for MLPs stems from the strong correlations in the posterior weight distribution, that is, the great majority of the candidate steps generated in the random walk will be rejected as they lead to a decrease in $p(\mathbf{w} | D)$ (Bishop 1995). The Hybrid Monte Carlo algorithm (Duane et al. 1987) reduces the random walk behaviour by using gradient information, which, in the case of MLPs, can be readily calculated.

While the Hybrid Monte Carlo algorithm allows for the efficient sampling of parameters (i.e. weights and biases), the posterior distribution for α should also be determined. In this paper we use Neal's (1996) approach and Gibbs sampling (Geman and Geman 1984) for α 's. The posterior α can be calculated using

$$\begin{aligned} p(\alpha | \mathbf{w}) &\propto p(\mathbf{w} | \alpha)p(\alpha) \\ &\propto \alpha^{m/2} \exp\left(-\frac{\alpha}{2} \sum_{i=1}^m w_i^2\right) \alpha^{a/2-1} \exp\left(-\frac{\alpha a}{2\mu}\right) \\ &\propto \alpha^{(m+a)/2-1} \exp\left(-\frac{\alpha}{2} \left(\frac{a}{\mu} + \sum_{i=1}^m w_i^2\right)\right), \end{aligned} \quad (9)$$

where $p(\mathbf{w} | \alpha)$ and $p(\alpha)$ are obtained from (5) and (6), respectively. Note that the posterior is again a gamma distribution, and that it is straightforward to calculate the shape and mean of the posterior distribution in terms of the old values.

Empirical Results

In this section, mineral potential maps produced using the Bayesian approach will be compared with maps produced using conventional gradient descent methods.

Study Region

The study region used is in the vicinity of Castlemaine, Australia, and extends from a Northwest corner with coordinates 251,250 mE, 5,895,250 mN, to a Southeast corner with coordinates 258,250 mE, 5,885,000 mN, where all specified coordinates

are Northings/Eastings referenced according to AMG Zone 55 AGD 66. Input data consisted of 16 input layers: three based on magnetics (magnetic field intensity, first derivative of magnetic field intensity, and automatic gain control filtered magnetics); five layers based on radiometrics (Th, U, K, TotalCount, K/Th); seven based on geochemistry (Au, As, Cu, Mo, Pb, W, Zn) and distance to closest fault. The study region was represented by a rectangular grid of 141 by 206 cells with resolution 50 m by 50 m. The number of documented gold deposits in the study area is 148.

Gradient Descent Training with Early Stopping

Results of applying gradient descent training are first provided. The algorithm used was the Scaled Conjugate Gradients algorithm (Møller 1993). Overfitting prevention was implemented with early stopping using the validation procedure described in Skabar (2005a). Throughout, we use a hidden layer consisting of 6 hidden layer units, which we know from previous work to be sufficient to adequately represent the target function.

The degree of variability between maps resulting from different training/validation set partitions was measured by generating 10 maps, each corresponding to a different randomly selected partition. In each case, the validation set consisted of 25% of the known mineralized pixels. As training progressed, the likelihood over these holdout examples was monitored, and the stopping point was chosen as the point at which this likelihood was a maximum. The MLP was then used to produce a mineral potential map. For illustrative purposes, two of the 10 maps are shown in Fig. 1, together with the training deposits (points) and validation deposits (crosses). Note the clearly discernable differences between the two maps.

An objective measure of the similarity between the two maps is provided by the Pearson correlation coefficient

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

where x and y are the pixel values (i.e. the conditional probabilities of mineralization) for each of the two maps, $\bar{x}$ and $\bar{y}$ are their respective means, and n is the number of pixels in the map. Table 1 shows the Pearson correlation coefficient between all pairwise combinations of the 10 maps produced from different training/validation set partitions. The mean correlation is 0.609 with the standard deviation of 0.109, reflecting a moderately strong positive correlation. Note that the maps in Fig. 1A and B correspond respectively to Map 1 and Map 8 from Table 1. These maps were chosen because their correlation is closest to the average over all pairwise combinations.

The variability observed above was due to the training/validation set partition. In a second set of experiments, we fixed the training/validation partition and produced further 10 maps, each corresponding to a different random weight initialization. In this case, the mean and standard deviation of the Pearson correlation coefficient for pairwise combinations of the maps were 0.608 and 0.103, respectively.

We measured how well the maps reflect the location of training and validation deposits by ranking pixels according to their conditional probability of mineralization,

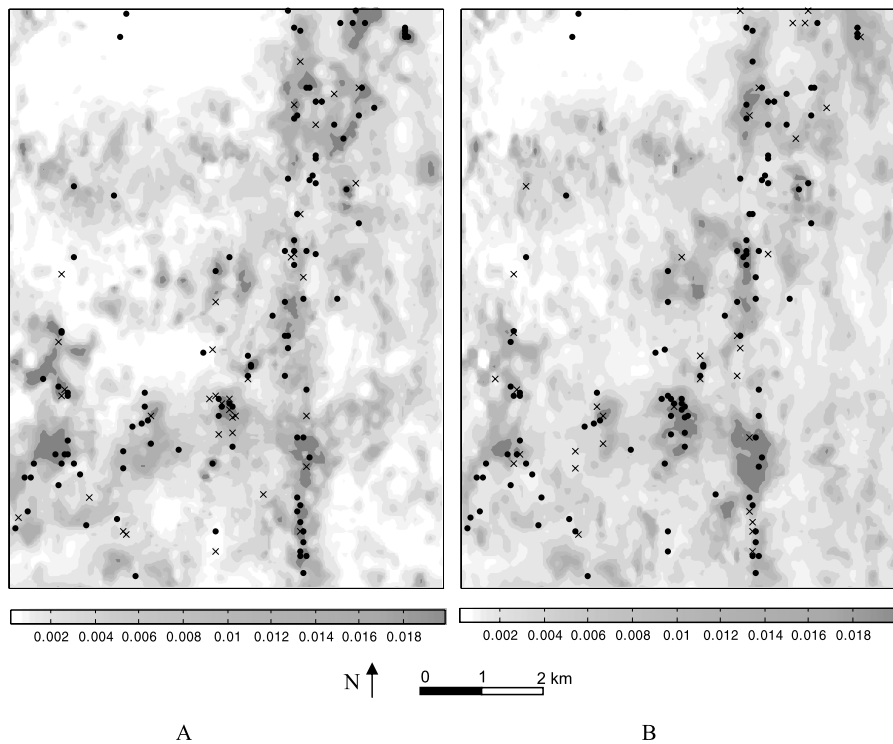


Fig. 1 Mineral potential maps produced from different training/validation set partitions using gradient descent. *Points* indicate the location of training deposits; *crosses* indicate location of validation deposits. Values on *color bar* are conditional probabilities of mineralization. The prior probability of mineralization is 0.0038. Maps extend from a Northwest corner with coordinates 251,250 mE, 5,895,250 mN, to a Southeast corner with coordinates 258,250 mE, 5,885,000 mN, where the specified coordinates are Northings/Eastings referenced according to AMG Zone 55 AGD 66

Table 1 Pearson Correlation Coefficient between pairwise combinations of 10 maps produced from different training/validation set partitions. Mean correlation is 0.609; standard deviation is 0.109

	Map 2	Map 3	Map 4	Map 5	Map 6	Map 7	Map 8	Map 9	Map 10
Map 1	0.571	0.423	0.669	0.586	0.647	0.557	0.604	0.641	0.685
Map 2	–	0.412	0.549	0.650	0.563	0.648	0.653	0.640	0.685
Map 3	–	–	0.483	0.419	0.690	0.404	0.754	0.757	0.489
Map 4	–	–	–	0.562	0.659	0.539	0.625	0.610	0.637
Map 5	–	–	–	–	0.504	0.671	0.589	0.549	0.551
Map 6	–	–	–	–	–	0.482	0.862	0.832	0.652
Map 7	–	–	–	–	–	–	0.652	0.521	0.524
Map 8	–	–	–	–	–	–	–	0.871	0.674
Map 9	–	–	–	–	–	–	–	–	0.680

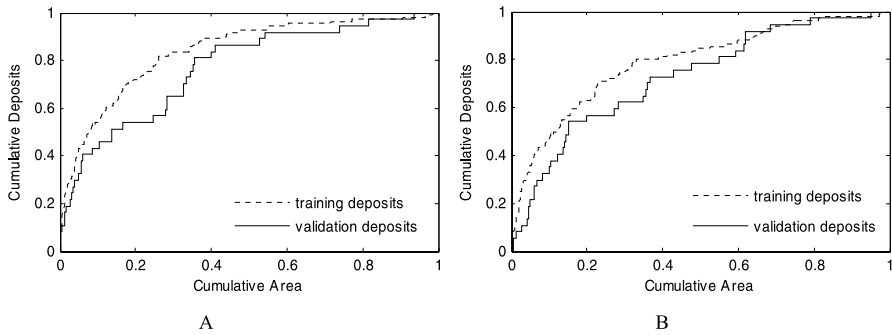


Fig. 2 Predictivity curves corresponding to mineral potential maps from Fig. 1

Table 2 Area under training deposit and validation deposit curves, averaged over maps produced from 10 randomly selected training/validation set partitions using gradient descent approach

	Mean	Standard deviation
Training deposits	0.8157	0.0485
Validation deposits	0.7514	0.0518

and plotting the proportion of observed deposits (both training and validation) against cumulative area, progressing from pixels of high potential to pixels of low potential. Figure 2 shows the plots corresponding to the two maps in Fig. 1. The quantity of interest is the area under the curves. The mean and standard deviation of the area under the curves for the first set of 10 maps is given in Table 2. As expected, the maps are better at reflecting the presence of training deposits than validation deposits. This is because the training deposits were used for weight optimization, whereas the validation deposits were used only for monitoring training performance. It is important to note that predicting validation deposits is not a measure of the ability to predict test deposits. In general, a map will not predict test deposits as well as it predicts validation deposits. A test set was not used because we were primarily interested in obtaining a reasonable measure of the degree of variability in the maps. Using a test set would have made fewer examples available for training and validation, thus increasing the variability.

Bayesian Learning

Results of the application of the Bayesian learning approach are provided in this section. The number of hidden units was again six, and the prior distribution for the hyperparameter α (i.e. the inverse square variance for the weight prior) was set to a gamma distribution with a mean of 0.1 and shape parameter of 1. This selection of values was guided by our knowledge of appropriate weight regularization coefficient values as observed in previous work (Skabar 2005a), and the known equivalence between this coefficient and α . For Monte Carlo sampling, a burn-in period of 2000 samples was used. The purpose of the burn-in period was to allow the sampling process to converge to the correct distribution. Following the burn-in period,

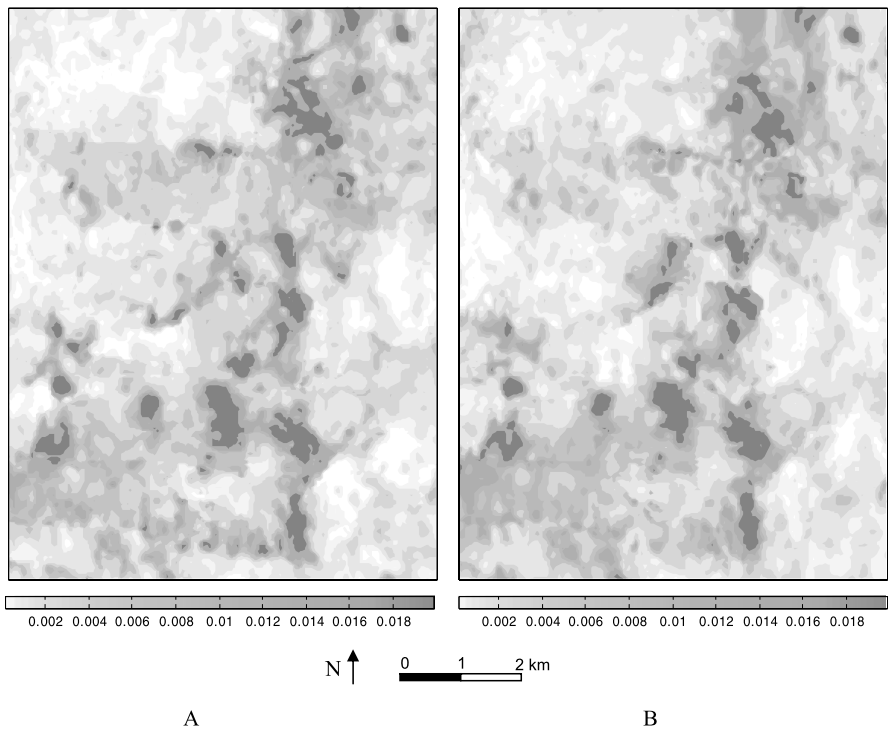


Fig. 3 Mineral potential maps produced using Bayesian approach. Pearson Correlation Coefficient between maps is 0.796. Maps **A** and **B** correspond to Maps 3 and 10 respectively from Table 3

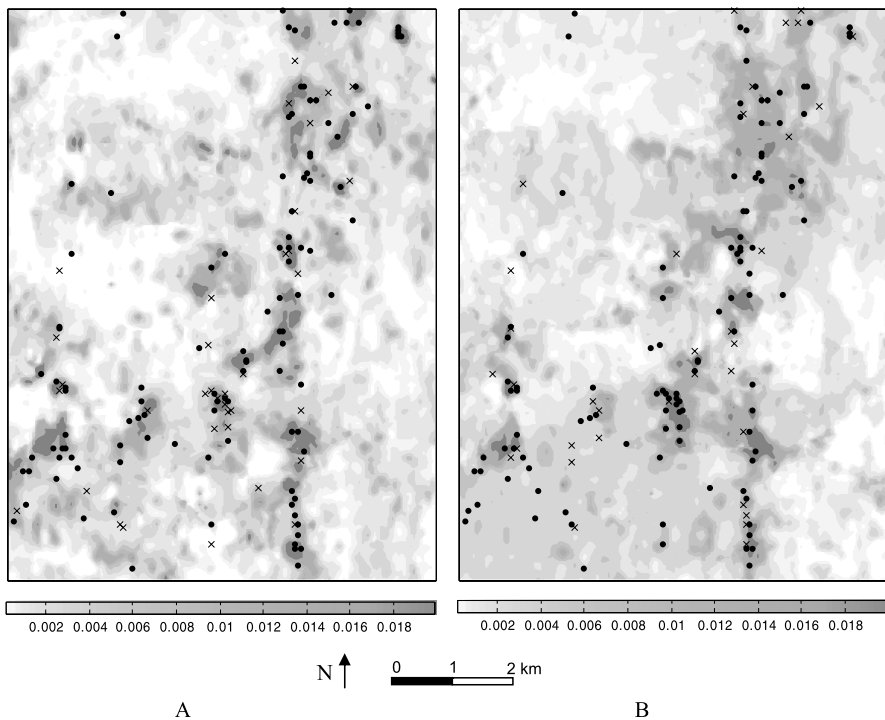
Table 3 Pearson Correlation Coefficient between pairwise combinations of 10 maps produced from separate runs of Bayesian approach. Mean correlation is 0.798; standard deviation is 0.063

	Map 2	Map 3	Map 4	Map 5	Map 6	Map 7	Map 8	Map 9	Map 10
Map 1	0.824	0.875	0.799	0.809	0.820	0.847	0.807	0.782	0.734
Map 2	–	0.863	0.864	0.881	0.821	0.832	0.861	0.727	0.802
Map 3	–	–	0.783	0.821	0.873	0.927	0.813	0.836	0.796
Map 4	–	–	–	0.946	0.697	0.760	0.848	0.657	0.791
Map 5	–	–	–	–	0.732	0.783	0.850	0.696	0.794
Map 6	–	–	–	–	–	0.803	0.764	0.740	0.735
Map 7	–	–	–	–	–	–	0.801	0.794	0.771
Map 8	–	–	–	–	–	–	–	0.698	0.793
Map 9	–	–	–	–	–	–	–	–	0.664

every tenth sample was collected and stored, until a total of 200 samples had been obtained. These 200 samples were then used to predict probabilities for each pixel in the study region; for each pixel, the mean was calculated.

Table 4 Area under training deposit and test deposit curves averaged over 10 maps produced from separate runs of Bayesian approach

	Mean	Standard deviation
Training deposits	0.8914	0.0151
Test deposits	0.7504	0.0528

**Fig. 4** Mineral potential maps produced from different training/test set partitions using Bayesian approach. Points indicate the location of training deposits; crosses indicate the location of test deposits. Dataset partitions are the same as those used in the maps of Fig. 1

The above procedure was applied 10 times, resulting in 10 maps. Two of these maps are shown in Fig. 3; notice the very close qualitative similarity between the maps. The correlation coefficients for each pairwise combination of the 10 maps are shown in Table 3. The average correlation was 0.798 with the standard deviation of 0.063. These values compare very favorably with those for gradient descent training and demonstrate that the Bayesian approach results in maps which display significantly less variability than those produced using gradient descent. However, the variation in the mean value for a pixel is inversely related to the number of samples collected, and thus the variability between maps could in principle be further reduced by simply using more samples for the calculation.

The Bayesian approach does not use a validation set, and the maps described above were produced using all examples for training. We can measure the ability of a map to predict the presence of test deposits by dividing the data into training/test sets, in an analogous way in which training/validation sets were created for the gradient descent method. That is, we use all data for weight optimization, but hide from the training algorithm the fact that some of the examples are mineralized. We can then measure the performance of the map in predicting these holdout test examples. We generated 10 maps in this way, each map corresponding to a different training/test set partition. The average area under the prediction curves is shown in Table 4. Note that the mean area under the test deposit curve (0.750) is not significantly different from the mean area under the validation deposit curve for gradient-descent training (0.751). This is an inspiring result, because, in general, performance on test examples will be below that of validation examples. This suggests that the expected test accuracy for the Bayesian approach is greater than that for the early-stopping approach. For illustrative purposes, two of the 10 maps produced using different training/test set partitions are shown in Fig. 4. These correspond to the same dataset partitions used in Fig. 1.

Conclusions

Reports on the application of mineral potential mapping techniques usually evaluate the success of a technique on the basis of how accurately the map predicts the presence of test deposits. An issue that is usually overlooked concerns the replicability of results; that is, ‘How much variability can we expect between two maps produced from the same data using the same technique?’ It is discomforting that the final map should depend to any significant extent on factors such as the particular choice of starting weights, or the training/validation set partition. Nevertheless, this phenomenon does occur when using gradient descent methods, particularly in high-dimensional, sparsely populated input spaces such as those typically used on mineral potential mapping tasks.

The advantage of the Bayesian approach is that it is integrative: it does not make predictions using a single weight vector. Rather, it performs a weighted sum of the predictions over all possible weight vectors, where the weighting coefficient for a particular weight vector depends on the posterior weight distribution. This prevents overfitting without having to use a validation procedure, and thereby eliminates much of the variability that is observed between maps produced using gradient descent methods. In principle, the variability between maps produced using the Bayesian approach is reduced to zero in the limit as the number of samples approaches infinity.

Our results suggest that the Bayesian approach is also very likely to be better than gradient descent methods at predicting the presence of unknown deposits. Once again, the best explanation for this is the integrative nature of the Bayesian approach: gradient descent methods are highly susceptible to noise in the training data, resulting in parameter values which may be sub-optimal, thus leading to inferior prediction performance. Bayesian methods avoid these pitfalls by averaging over many weight vectors.

We have assumed a fixed number of hidden layer units. The performance of gradient descent techniques can be very sensitive to this choice, and in general, the optimal number of hidden units should also be determined using validation (Skabar 2005a). Neal (1996) argues that the Bayesian learning approach will not overfit the training data, even as the number of hidden layer units approaches infinity. Thus, providing that the number of hidden layer units is sufficient to model the target function, it doesn't matter if this number is exceeded; however, computational requirements demand that the number of hidden layer units not be too large.

Aside from the number of hidden layer units, the only other item which must be pre-specified is the prior weight distribution. In our scheme, this distribution is controlled through α , the inverse squared variance of the Gaussian distribution of the sum of squared weights. Note, however, that we do not need to set an actual value for α , but only for the parameters of the gamma distribution that models the distribution of α . While these (gamma distribution) parameters will certainly influence the posterior weight distribution, the influence will not be as great as it would be through specifying α directly. This is supported by the results of Husmeier et al. (1998), who investigated the influence of various priors on generalization performance, and concluded that, overall, the Bayesian method is relatively robust with respect to the choice of prior distribution (Husmeier et al. 1998).

The purpose of this paper has been to compare the performance of MLPs trained using Bayesian methods with that of gradient descent trained MLPs. Although we have not directly compared these results with other approaches, we know from previous work (Skabar 2005a) that gradient descent trained MLPs are able to achieve better predictive performance than logistic regression, and predictive performance at least equal to that of generative approaches based on explicitly estimating distributions. Thus, the Bayesian approach provides yet a further refinement in the application of neural networks to mineral potential mapping.

In conclusion, Bayesian learning methods provide a principled, probabilistic approach that takes out much of the guesswork in applying neural network models. In the context of mineral potential mapping, they lead to maps which display significantly less variability than those due to gradient descent methods. Their integrative nature also makes them more reliable at predicting the presence of unknown deposits.

References

- Bishop C (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
- Brown WM, Gedeon TD, Groves DI, Barnes RG (2000) Artificial neural networks: a new method for mineral prospectivity mapping. *J Aust Earth Sci* 7(4):757–770
- Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst* 2:303–314
- Denison DGT, Holmes CC, Mallick BK, Smith AFM (2002) Bayesian methods for nonlinear classification and regression. Wiley, Chichester
- Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. *Phys Lett B* 195(2):216–222
- Geman S, Geman G (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
- Harris D, Pan G (1999) Mineral favorability mapping: a comparison of artificial neural networks, logistic regression, and discriminant analysis. *Nat Resour Res* 8(2):93–109
- Hornik K, Stinchcombe M, White H (1989) Multilayer feed-forward networks are universal approximators. *Neural Netw* 2(5):359–366

- Husmeier D, Penny WD, Roberts SJ (1998) Empirical evaluation of Bayesian sampling for neural classifiers. In: Niklasson L, Boden M, Ziemke T (eds) Proceedings of international conference on artificial neural networks (ICANN98). Perspectives in neural computing. Springer, pp. 323–328
- MacKay DJC (1992) A practical Bayesian framework for backpropagation networks. *Neural Comput* 4(3):448–472
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21(6):1087–1092
- Møller M (1993) A scaled conjugate gradient algorithm for fast supervised. *Neural Netw* 6(4):523–533
- Neal RM (1996) Bayesian for neural networks. Lecture notes in statistics, vol 118. Springer, New York
- Porwal A, Carranza EJ, Hale M (2003) Artificial neural networks for mineral-potential mapping: a case study from the Aravalli Province, Western India. *Nat Resour Res* 12(3):155–171
- Skabar A (2005a) Mapping mineralization probabilities using multilayer perceptrons. *Nat Resour Res* 14(2):109–123
- Skabar A (2005b) Application of Bayesian MLP techniques to predicting mineralization potential from geoscientific data. In: Duch W, Kacprzyk J, Oja E, Zadrozny S (eds) Proceedings of international conference on artificial neural networks (ICANN 2005). Lecture notes in computer science, vol 3697. Springer, New York, pp 963–968