Can. J. Remote Sensing, Vol. 31, No. 4, pp. 269–273, 2005

Research Note / Note de recherche

On the use of consensus algorithms to address variability in the results of neural network classifications: preliminary tests involving two northern study areas

David W. Leverington and Wooil M. Moon

Abstract. In past work it has been recognized that variations in parameters such as learning rate, momentum, and network architecture can influence the results in neural network classifications of satellite images. New tests suggest that variation in the results of neural network classifications, caused solely by differences in weight initializations, can also be substantial. This issue has the potential to limit the applicability of neural networks in remote sensing classifications. The negative effects of variation in neural network results can potentially be reduced or eliminated through application of consensus algorithms in which the outputs of multiple neural network classifications are combined. Research results presented here were based on training and test data with low sample sizes for many classes and, accordingly, the results must be interpreted with caution. Early results using majority-vote and evidential-reasoning consensus algorithms, however, suggest that near-optimum neural network classification accuracies can be achieved through application of these algorithms.

Résumé. Il a été établi, dans des travaux de recherche antérieurs, que des variations dans les paramètres tels que le taux d'apprentissage, le momentum et l'architecture du réseau peuvent influencer les résultats de classification dans les classifications d'images par réseau de neurones. De nouveaux tests suggèrent que la variation dans les résultats de classification par réseau de neurones causée uniquement par des différences dans les initialisations des poids peut également être substantielle. Cette problématique pourrait limiter l'applicabilité des réseaux de neurones dans les classifications en télédétection. Les effets négatifs de la variation dans les résultats des réseaux de neurones peuvent potentiellement être réduits ou éliminés par le biais de l'application d'algorithmes de consensus dans lesquels les extrants de classifications multiples par réseau de neurones sont combinés. Les résultats de recherche présentés ici sont basés sur des données d'entraînement et d'expérimentation caractérisées par des échantillons de faible dimension pour plusieurs classes et, conséquemment, les résultats doivent être interprétés avec prudence. Toutefois, les résultats préliminaires utilisant les algorithmes de consensus basés sur le vote majoritaire et le raisonnement par évidence indiquent qu'en appliquant ces derniers, on peut atteindre des précisions quasi optimales de classification par réseau de neurones.
[Traduit par la Rédaction]

Introduction

An important problem associated with the use of feed-forward, back-propagation neural networks, the most widely used neural networks in image classification, is the potential for substantial variability in final classification results between individual classifications. Variations in network parameters such as learning rate, momentum, and the nature of hidden nodes can all impact classification results. Furthermore, even when all other factors (including training data, learning algorithm, and network topology) are kept constant, the random initialization of network weights prior to each execution of a neural network training algorithm can cause final classification results to vary from execution to execution. This weight issue alone can reduce the utility of neural networks in image classification by necessitating the generation and manual evaluation of multiple sets of results to ensure the production of the most satisfactory result possible. Consensus algorithms, which combine multiple sets of neural network outputs to produce new results, have the potential to address this issue by automatically generating optimum or near-optimum classification results.

Using two northern study areas as a basis, this preliminary study undertook to determine the magnitude of variability that can be involved in neural network classifications of satellite images and to test two consensus algorithms (which combined

Received 11 March 2004. Accepted 1 August 2005.

D.W. Leverington[1] and W.M. Moon. Department of Geological Sciences, University of Manitoba, Winnipeg, MB R3T 2N2, Canada.

[1]Corresponding author. Present address: Department of Geosciences, Texas Tech University, Box 41053, Lubbock, TX 79409-1053, USA (e-mail: david.leverington@ttu.edu).

© 2005 CASI

sets of neural network results using majority-vote and evidential-reasoning routines) in the mitigation of this issue.

Consensus algorithms

Consensus theory involves the search for consensus among a group of "experts" to improve the outcome of a decision-making process (e.g., see Benediktsson and Swain, 1992). Consensus algorithms (e.g., Battiti and Colla, 1994; Ji and Ma, 1997; Benediktsson et al., 1997a; Khotanzad et al., 2000) range in complexity from simple majority-vote rules (e.g., Hansen and Salamon, 1990; Jiminez, 1999) to more complex rules that involve, for example, the combination of neural network output activations using other neural networks (e.g., Benediktsson et al., 1997b). In this research, two consensus algorithms were explored as possible solutions to the issue of variability in neural network classifications caused by differences in initial network weights during training.

The majority-vote algorithm was chosen for evaluation as a consensus procedure on the basis of its simplicity and its ease in software implementation; this algorithm simply combined the final thematic results of neural network classifications by majority vote (**Figure 1**). Though more complex, an evidential-reasoning algorithm was also tested here as a consensus procedure on the basis of its promising performance in past neural network work (e.g., Rogova, 1994). The evidential-reasoning algorithm is based on the derivation of a mass of evidence for a given set of all-encompassing class propositions and the combination of this evidence through the technique of "Dempster's orthogonal sum" (see full description in, e.g., Richards, 1999); the class label whose combined evidence is greatest is assigned to the pixel in question. The evidential-reasoning consensus algorithm used in this research (programmed in C by the first author on the basis of Richards (1999)) treated the activations of neural network output nodes directly as measures of evidence, under the assumption that the relative magnitudes of such activations are roughly proportional to corresponding posterior probabilities (see, e.g., Ruck et al., 1990; Wan, 1990).

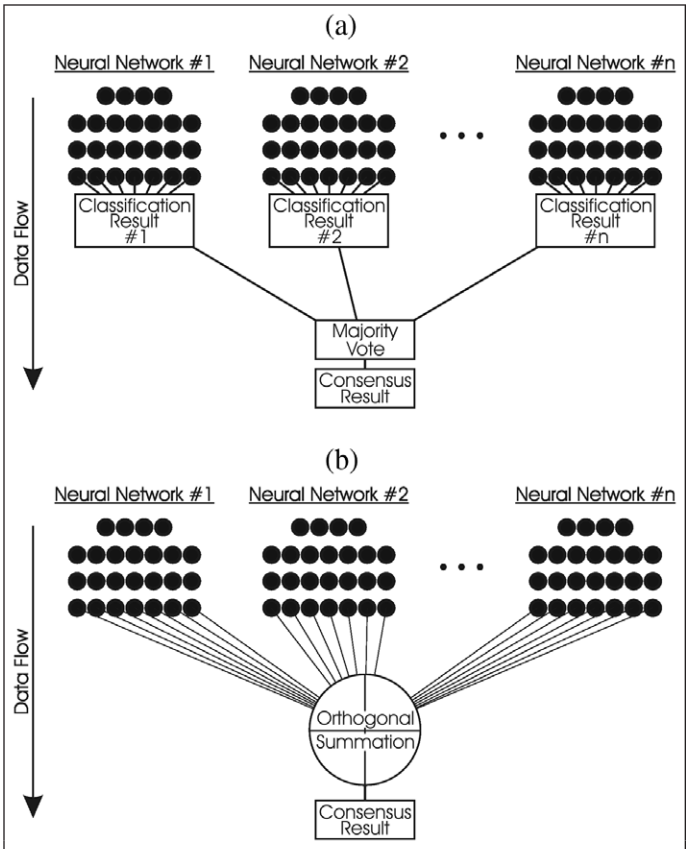

Figure 1. Schematic overview of the two consensus algorithms used in this research: (a) majority vote, and (b) orthogonal summation (evidential reasoning). The majority-vote consensus algorithm chooses the class label most frequently selected by individual neural networks. The evidential-reasoning consensus algorithm assigns class labels based on orthogonal summation of neural network output activations.

Neural network algorithm

The neural network software used in this study was programmed in C by the first author, with parameters and network architectures based on established neural network principles (Gallant, 1993; Bishop, 1995). Specifically, individual neural network executions were generated by a standard feed-forward network that uses back-propagation to calculate derivatives of training error and to adjust weights to minimize error. The error function used by the network is the sum-of-squares error. The sigma nonlinearity in [0.0, 1.0] is used as the activation function for all hidden and output nodes. Settings for the learning rate (ε) and momentum (α) were 0.1 and 0.9, respectively. Initial weights were uniformly distributed in [–0.5, +0.5]. The network was configured for classification using an output layer in which a unique node is assigned to each class; for a given input pattern, the assigned class is that whose output node has the highest activation. During training, target activations for "correct" and "incorrect" nodes were 0.9 and 0.1, respectively. Two intermediate layers were used for all executions and were defined with as many nodes as the maximum of the number of nodes in the input and output layers.

Study areas and satellite images

Two study areas, one located on Melville Island, Nunavut, and another in the Cape Smith Belt of northern Quebec, were used as a basis for this study. The Melville Island study area is a 20 km by 20 km region located in the central portion of eastern Melville Island (Leverington, 2001). Surface materials at this study area are dominantly comprised of weathered and frost-shattered felsenmeer (Hodgson et al., 1984), the lithologies of which include clastics, carbonates, and gabbro (e.g., Harrison, 1995) (**Table 1**). The Cape Smith Belt study area is a 30 km by 20 km region located in northern Quebec (Leverington, 2001;

© 2005 CASI

Table 1. Surface classes of the Melville Island study area.

Class	Description	No. of pixels		
		Total	Training	Test
1	Canyon Fiord Formation (red-weathering calcareous sandstone)	86	56	30
2	Canyon Fiord Formation (yellow calcareous clastic fines)	20	13	7
3	Canyon Fiord Formation (pink clastic fines)	20	13	7
4	Assistance Formation mudstone	35	23	12
5	Sabine Bay Formation sandstone	90	59	31
6	Trold Fiord Formation sandstone	78	51	27
7	Bjorn Formation sandstone	32	21	11
8	Tingmisut Inlier dolomite (Beverly Inlet Formation)	74	48	26
9	Great Bear Cape Formation limestone	135	89	46
10	Degerböls Formation limestone	48	31	17
11	Gabbro	40	26	14
12	Green vegetation	41	27	14
13	Snow cover	35	23	12
14	Water	117	77	40

see also, e.g., St-Onge et al., 1999). Felsenmeer and bedrock exposure in the study area is extensive; exposed surface classes include (*i*) basalt, gabbro, and peridotite of the Watts Group; (*ii*) pelite of the Spartan Group; and (*iii*) basalt of the Chukotat Group (**Table 2**).

Landsat-5 thematic mapper (TM) images were used as a basis for the classifications performed in this study. The image for the Melville Island study area was acquired on 8 August 1994 (TM media number 409069), and the image for the Cape Smith Belt study area was acquired on 3 September 1997 (TM media number 972462). Data from four Landsat TM channels were used as input to the image classifications: channel 3 (0.63–0.69 μm), channel 4 (0.76–0.90 μm), channel 5 (1.55–1.75 μm), and channel 7 (2.08–2.35 μm). The selection of TM channels thus emphasised red to mid-infrared wavelengths, where variation between the reflectance properties of individual rock types is typically greatest (e.g., Drury, 1993).

Classification methodology

A total of 851 pixel locations in the Melville Island Landsat TM image were selected for use as training- and test-pixel locations for 14 different surface-cover classes (**Table 1**), and a total of 686 pixel locations in the Cape Smith Belt Landsat TM image were selected for use as training- and test-pixel locations for seven different surface-cover classes (**Table 2**). Six different training and test databases were used in this research to more clearly discern trends in classification results and reduce the effects of the small sample sizes of some classes. Each of the six databases for the Melville Island study area was generated by dividing (randomly and stratified by class) the 851 pixel locations into 557 training pixels and 294 test pixels (the ratio of training pixels to test pixels using this scheme is roughly 2:1). Each of the six databases for the Cape Smith Belt series was generated by dividing the 686 pixel locations into 449 training pixels and 237 test pixels, again based on random sampling, stratified by class.

Table 2. Surface classes of the Cape Smith Belt study area.

Class	Description	No. of pixels		
		Total	Training	Test
1	Peridotite	162	106	56
2	Pelite	70	46	24
3	Chukotat Group basalt	158	104	54
4	Watts Group basalt	112	73	39
5	Tonalite	36	23	13
6	Green vegetation	84	55	29
7	Water	64	42	22

For each set of classifications the following were generated: (*i*) one maximum likelihood classification (generated using the Geomatica 8.1 software package, PCI Geomatics Enterprises Inc., Richmond Hill, Ont.) to provide an informal classification benchmark; (*ii*) 10 individual neural network classifications, each based on a different random set of initial weights; (*iii*) one consensus classification based on a majority vote of the final results of the 10 individual neural network classifications (in the rare event of a tie, the class with the lowest numerical designation was arbitrarily assigned); and (*iv*) one consensus classification that combined output activations of the 10 individual neural network classifications by orthogonal summation (i.e., through application of the evidential-reasoning algorithm).

Classification results

Classification results for the Melville Island study area, expressed both as overall percentages of test pixels correctly labelled and as Kappa statistics, are summarized in **Table 3** (note that, because the training and test data of some classes were characterized by relatively small sample sizes, the Melville Island and Cape Smith Belt results, discussed in the following, are considered preliminary). Box plots of the ranges of overall percentages associated with each of the six sets of 10 individual neural network executions are given in **Figure 2**;

© 2005 CASI

Table 3. Summary of overall classification results for the Melville Island study area, given for six different training and test databases (sets 1–6).

Algorithm	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Avg.
NN-B	0.7755 (0.7506)	0.7755 (0.7522)	0.8027 (0.7810)	0.7857 (0.7620)	0.7721 (0.7475)	0.7789 (0.7550)	0.7817 (0.7581)
NN-M	0.7857 (0.7624)	0.7891 (0.7668)	0.7891 (0.7661)	0.7789 (0.7552)	0.7687 (0.7440)	0.7789 (0.7559)	0.7817 (0.7584)
NN-ER	0.7925 (0.7696)	0.7857 (0.7632)	0.7925 (0.7701)	0.7687 (0.7442)	0.7517 (0.7262)	0.7755 (0.7521)	0.7778 (0.7542)
ML	0.7653 (0.7411)	0.7347 (0.7087)	0.7313 (0.7037)	0.7211 (0.6933)	0.7415 (0.7157)	0.7585 (0.7344)	0.7421 (0.7162)

Note: Plots of variability in the overall results of individual neural network executions are given in **Figure 2**. Values are expressed as proportions of test pixels classified correctly and may be converted to percentages by multiplying by 100%. Values in parentheses are kappa (KHAT) statistic measures of the percentage of test pixels correctly labelled beyond that expected on chance alone. Classification algorithms are labelled as follows: ML (maximum likelihood result); NN-B (best individual neural network result of 10 neural network executions); NN-ER (evidential-reasoning consensus result); NN-M (majority-vote consensus neural network result).

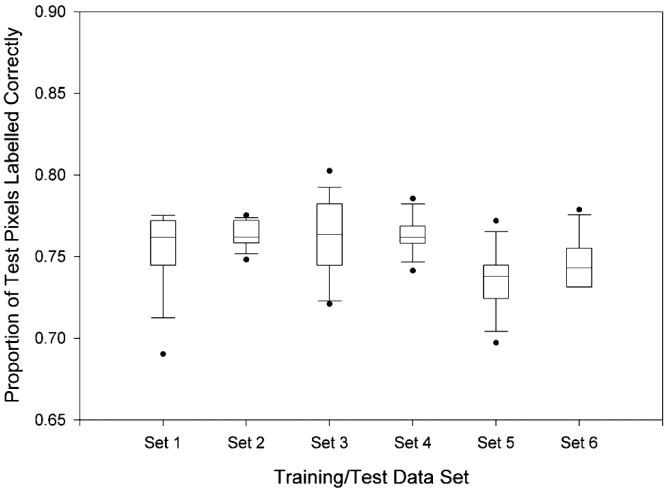

Figure 2. Box plots showing the variation in individual overall classification results produced by the neural network classifier for the Melville Island study area. The boxes give the 25th, 50th, and 75th percentiles, and the whiskers (vertical lines that end in a horizontal stroke) give the 5th, 10th, 90th, and 95th percentiles.

variation in individual overall neural network classification results was as great as ~8%. The effects of this wide range in results were mitigated by both the majority-vote and the evidential-reasoning consensus algorithms: both sets of consensus neural network results were typically within ~1.5% of the best individual neural network result. Neither consensus algorithm produced results substantially superior to the best individual neural network result, however, with average proportions of test pixels correctly labelled of ~78% (~4% higher than the maximum likelihood classifier).

Classification results for the Cape Smith Belt study area are summarized in **Table 4**, and box plots of the ranges of overall percentages associated with each of the six sets of 10 individual neural network executions are given in **Figure 3**; as with the Melville Island study area, variation in individual overall neural network classification results was as great as ~8%. The effects of this wide range in results were again mitigated by both the majority-vote and evidential-reasoning consensus algorithms, and both sets of consensus neural network results were typically within ~1.5% of the best individual neural network result. Again, neither consensus algorithm produced results substantially superior to the best individual neural network result, with average proportions of test pixels correctly labelled of ~89% (~4% higher than the maximum likelihood classifier).

Early conclusions

In this preliminary study, variations in individual overall neural network classification results, caused simply by differences in initial randomly set weight values, were as great as ~8%. Results tentatively suggest that variability in the nature of neural network results can be addressed through the automated generation of multiple neural network results and the subsequent application of either a majority-vote or evidential-reasoning consensus algorithm to these results. This method can effectively and automatically produce an optimum or near-optimum neural network result, without necessitating manual evaluation of multiple individual neural network classifications (though user assessment of the final classification remains necessary). Results obtained were based on training and test data with low sample sizes for many classes; accordingly, the results must be interpreted with caution. Future work will be conducted with larger and more diverse datasets to confirm the results presented here, determine more specifically the minimum number of individual classifications necessary for the technique to be effective, and determine how techniques such as the use of a variety of network architectures and learning algorithms (e.g., Rogova, 1994) or the use of mathematical transforms of network inputs (e.g., Benediktsson et al., 1997a) might produce consensus results that are superior to all individual results used as input. Future work will also examine the sensitivity of image classification results to variations in other important neural network parameters such as learning rate, momentum, and network architecture.

Acknowledgments

The helpful comments of two anonymous reviewers are appreciated. This work was supported in part by the Geological Society of America, the Northern Scientific Training Program (Department of Indian Affairs and Northern Development Canada), Falconbridge Ltd., and the Smithsonian Institution (D.W.L.), and by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (A-7400) (W.M.M.).

© 2005 CASI

Table 4. Summary of overall classification results for the Cape Smith Belt study area, given for six different training and test databases (sets 1–6).

Algorithm	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Avg.
NN-B	0.9156 (0.8981)	0.8776 (0.8517)	0.8987 (0.8775)	0.8903 (0.8678)	0.8861 (0.8627)	0.8945 (0.8727)	0.8938 (0.8718)
NN-M	0.9030 (0.8823)	0.8776 (0.8516)	0.8987 (0.8778)	0.8945 (0.8725)	0.8734 (0.8474)	0.8692 (0.8422)	0.8861 (0.8623)
NN-ER	0.9156 (0.8978)	0.8565 (0.8263)	0.8987 (0.8778)	0.8819 (0.8567)	0.8903 (0.8674)	0.8861 (0.8620)	0.8882 (0.8647)
ML	0.8734 (0.8484)	0.7890 (0.7485)	0.8523 (0.8244)	0.8692 (0.8435)	0.8608 (0.8334)	0.8397 (0.8084)	0.8474 (0.8178)

Note: Plots of variability in the overall results of individual neural network executions are given in **Figure 3**. Values are expressed as proportions of test pixels classified correctly and may be converted to percentages by multiplying by 100%. Values in parentheses are kappa (KHAT) statistic measures of the percentage of test pixels correctly labelled beyond that expected on chance alone. Classification algorithms are labelled as in **Table 3**.

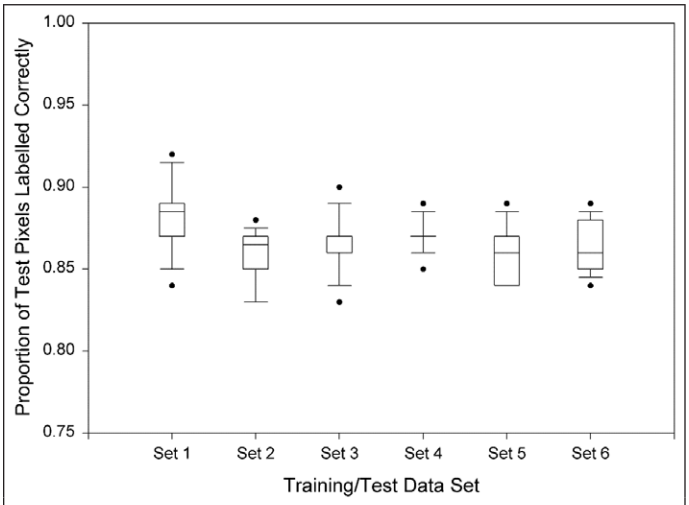

Figure 3. Box plots showing the variation in individual overall classification results produced by the neural network classifier for the Cape Smith Belt study area. The boxes give the 25th, 50th, and 75th percentiles, and the whiskers (vertical lines that end in a horizontal stroke) give the 5th, 10th, 90th, and 95th percentiles.

References

Battiti, R., and Colla, A.M. 1994. Democracy in neural nets: voting schemes for classification. *Neural Networks*, Vol. 7, pp. 691–707.

Benediktsson, J.A., and Swain, P.H. 1992. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 22, pp. 688–704.

Benediktsson, J.A., Sveinsson, J.R., Ersoy, O.K., and Swain, P.H. 1997a. Parallel consensual neural networks. *IEEE Transactions on Neural Networks*, Vol. 8, pp. 54–64.

Benediktsson, J.A., Sveinsson, J.R., Ingimundarson, J.I., Sigurdsson, H.S., and Ersoy, O.K. 1997b. Multistage classifiers optimized by neural networks and genetic algorithms. *Nonlinear Analysis, Theory, Methods, and Applications*, Vol. 30, pp. 1323–1334.

Bishop, C.M. 1995. *Neural networks for pattern recognition*. 2nd ed. Oxford University Press, Oxford, UK.

Drury, S.A. 1993. *Image interpretation in geology*. Chapman and Hall, New York.

Gallant, S.I. 1993. *Neural network learning and expert systems*. MIT Press, Cambridge, Mass.

Hansen, L.K., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, pp. 993–1001.

Harrison, J.C. 1995. *Melville Island's salt-based fold belt, arctic Canada*. Geological Survey of Canada, Bulletin 472.

Hodgson, D.A., Vincent, J.S., and Fyles, J.G. 1984. *Quaternary geology of central Melville Island, Northwest Territories*. Geological Survey of Canada, Paper 83-16.

Ji, C., and Ma, S. 1997. Combinations of weak classifiers. *IEEE Transactions on Neural Networks*, Vol. 8, pp. 32–42.

Jiminez, L.O. 1999. Classification of hyperdimensional data based on feature and decision approaches using projection pursuit, majority voting, and neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, pp. 1360–1366.

Khotanzad, A., Elragal, H., and Lu, T.-L. 2000. Combination of artificial neural-network forecasters for prediction of natural gas consumption. *IEEE Transactions on Neural Networks*, Vol. 11, pp. 464–473.

Leverington, D.W. 2001. *Discriminating lithology in arctic environments from earth orbit: an evaluation of satellite imagery and classification algorithms*. Ph.D. thesis, Department of Geological Sciences, University of Manitoba, Winnipeg, Man.

Richards, J.A. 1999. *Remote sensing digital image analysis*. 3rd ed. Springer-Verlag, New York.

Rogova, G. 1994. Combining the results of several neural network classifiers. *Neural Networks*, Vol. 7, pp. 777–781.

Ruck, D.W., Rogers, S.K., Kabrisky, M., Oxley, M.E., and Suter, B.W. 1990. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, Vol. 1, pp. 296–298.

St-Onge, M.R., Lucas, S.B., Scott, D.J., and Wodicka, N. 1999. Upper and lower plate juxtaposition, deformation and metamorphism during crustal convergence, Trans-Hudson Orogen (Quebec–Baffin segment), Canada. *Precambrian Research*, Vol. 93, pp. 27–49.

Wan, E.A. 1990. Neural network classification: a Bayesian interpretation. *IEEE Transactions on Neural Networks*, Vol. 1, pp. 303–304.

© 2005 CASI