Appears in *Proceedings, Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms Workshop,* Thirteenth National Conference on Artificial Intelligence, Portland, OR: AAAI Press (1996).

Human Expert-Level Performance on a Scientific Image Analysis Task by a System Using Combined Artificial Neural Networks

Kevin J. Cherkauer

Department of Computer Sciences
University of Wisconsin-Madison
1210 West Dayton Street
Madison, WI 53706, USA
cherkauer@cs.wisc.edu

Abstract

This paper presents the PLANNETT system, which combines artificial neural networks to achieve expert-level accuracy on the difficult scientific task of recognizing volcanos in radar images of the surface of the planet Venus. PLANNETT uses ANNs that vary along two dimensions: the set of input features used to train and the number of hidden units. The ANNs are combined simply by averaging their output activations. When PLANNETT is used as the classification module of a three-stage image analysis system called JARTOOL, the end-to-end accuracy (sensitivity and specificity) is as good as that of a human planetary geologist on a four-image test suite. JARTOOL-PLANNETT also achieves the best algorithmic accuracy on these images to date.

Introduction

Between 1991 and 1994, the Magellan space probe collected more than 30,000 synthetic aperture radar (SAR) images of the surface of the planet Venus, a greater amount of data than all previous planetary missions combined (Smyth *et al.* 1995). To analyze this data, researchers at NASA's Jet Propulsion Laboratory (JPL) have developed an AI system, JARTOOL (Burl *et al.* 1994), to automatically detect and catalog the approximately one million small volcanos estimated to be visible in the images, as a prelude to more advanced planetary geology studies.

The accuracy of the current JARTOOL system is still substantially below that of human scientists as measured by its sensitivity (ability to detect true volcanos) and specificity (ability to avoid detecting nonvolcanos). This motivates the current work, which introduces the PLANNETT system (Person-Level Artificial Neural Networks for ExtraTerrestrial Terrain classification) in an attempt to improve JARTOOL's accuracy to the level of a human expert. PLANNETT is a machine learning (ML) system that combines multiple artificial neural networks (ANNs) to improve volcano classification performance. When PLANNETT is substituted for JARTOOL's current classification module, the resulting end-to-end accuracy of the JARTOOL-PLANNETT system reaches that of a human planetary geologist on a four-image test suite.

As background, we briefly discuss the establishment of ground truth for this problem, needed for performance comparisons, and give an overview of the JARTOOL system. Then we present the PLANNETT system, and compare the new, combined JARTOOL-PLANNETT system to the current JARTOOL (called here JARTOOL-GAUSS) and to the performance of two human planetary geologists at detecting volcanos in a test suite of four images.

Establishing Ground Truth

Because we cannot verify which objects in the images are truly volcanos, we must rely on human experts to establish a ground truth labeling. Ground truth is needed as a reference both for training ML volcano classifiers and for comparing the performance of various algorithms and humans. Because there is a wide variation in the labelings created by individual experts, the ground truth used for this study was the consensus of two planetary geologists who discussed and labeled the images together (Fayyad & Smyth 1993). This is the ground truth labeling used for most of the JARTOOL work at JPL. It consists of 163 volcanos in four sample Magellan-SAR images. (Smyth *et al.* (1995) describe an alternative, algorithmic approach to estimating ground truth from the individual labelings of many experts.)

The Current JARtool-Gauss System

JPL's current JARTOOL system (Burl *et al.* 1994), called here JARTOOL-GAUSS, consists of three modules

- Focus of Attention
- Feature Measurement
- Classification

The focus of attention (FOA) module uses a matched filter to quickly scan the entire image database for regions possibly containing volcanos. The goals of this module are speed and sensitivity. The system attempts to find as many of the true volcanos as possible, but

Table 1: Summary of the eight input representations PLANNETT uses. GAUSS uses only representation 4.

Rep. Name	Feats	Features Included
1. PCC6-HighRes	6	First six PCCs at full resolution
2. PCC6-MedRes	6	First six PCCs at half resolution
3. PCC6-LowRes	6	First six PCCs at quarter resolution
4. PCC6-MedRes-Petal	7	PCC6-MedRes, max petal filter
5. PCC6-AllRes-Petal	19	PCC6-HighRes, PCC6-MedRes, PCC6-LowRes, max petal filter
6. FFT-Feats	25	Avg. energies collected along 18 radii and 7 concentric rings of FFT image
7. All-But-FFT	94	All-Feats minus FFT-Feats
8. All-Feats	119	12 PCCs and 12 RMS reconstruction errors at full, half, and quarter res.; FFT-Feats; features of dark, medium, and bright regions; max petal filter

may generate many false alarms (regions with no volcanos) in the process. That is, FOA has high sensitivity but low specificity.

The FOA module generates a list of candidate image patches to be examined more carefully for volcanos. This list is passed to the feature measurement module, which measures a set of 119 continuous-valued descriptive features from each patch. The features attempt to capture the relevant information needed to distinguish volcanos from false alarms. The complete set of 119 available features consist of the first 12 principal component coefficients (PCCs) from a singular value decomposition of the true volcanos in the training set and the corresponding 12 RMS reconstruction errors, each at three different image patch resolutions (72 features); average energies collected along 18 radii and 7 concentric rings of the Fast Fourier Transform of each patch (25 features); 7 features of the dark, medium, and bright regions of each patch, segmented using the algorithm of (Pappas 1992) (21 features); and the maximum response of an orientation-dependent "petal" filter that responds most strongly to elongated, linear objects (1 feature). (A high petal response is thus evidence against a volcano.)

The classification module then uses the measured features to decide which candidate patches actually contain volcanos. It uses ML techniques and a set of labeled training examples to learn to distinguish volcanos from false alarms, and then classifies each candidate according to what it has learned.

JARTOOL-GAUSS uses a simple Gaussian classifier (Burl *et al.* 1994), called here GAUSS, as its classification module. GAUSS examines only seven input features: the first six PCCs of each candidate patch at half resolution and the maximum petal response (representation PCC6-MedRes-Petal in Table 1).[1] It fits a seven-dimensional Gaussian to the training examples of each class (*volcano* and *nonvolcano*) and uses Bayes' rule to estimate the posterior probability that a given testing example is a volcano. GAUSS currently does not use the other 112 available features.

[1] In (Burl *et al.* 1994), GAUSS did not use the petal feature, but further work has shown that including it slightly improves accuracy.

The New JARtool-Plannett System

PLANNETT is an ML system that replaces JARTOOL-GAUSS's current GAUSS classification module to produce a new algorithm, JARTOOL-PLANNETT. We developed PLANNETT in an attempt to take advantage of the information contained in all 119 available input features. PLANNETT consists of a set of 32 feed-forward ANNs individually trained to distinguish volcanos from nonvolcanos using back propagation (Rumelhart, Hinton, & Williams 1986), and combined by averaging their output activations. The ANNs contain an input layer, two output nodes (*volcano* and *nonvolcano*), and an optional hidden layer. The final classification is made by thresholding the difference between the averaged activations of the volcano and nonvolcano output units of all 32 ANNs. The averaging used to combine the ANNs is unweighted and thus takes no account of the differing classification accuracies of individual ANNs. Despite its simplicity, this style of combining outputs has been shown to produce a good composite model (Clemen 1989; Krogh & Vedelsby 1995). More sophisticated combining techniques are described in, e.g. (Wolpert 1992; Krogh & Vedelsby 1995; Tresp & Taniguchi 1995).

The ANNs vary on two dimensions: the subset of input features they are trained on and the number of hidden units they contain. PLANNETT trains ANNs using eight different subsets of the 119 available continuous-valued input features. Each of these subsets, or *representations*, was hand selected as a potentially sensible grouping of related features. The representations are summarized in Table 1. Representations 1–4 in the table concentrate on the first six PCCs at different resolutions and the petal filter because these features work well with the Gaussian classifier. The larger representations 5–8 in Table 1 were developed specifically for the ANNs, because the ANNs tended to become more accurate as more features were added.

For each of the eight representations, PLANNETT trains four ANNs that contain, respectively, 0, 5, 10, and 20 hidden units in one layer. The parameters of ANN training are summarized in Table 2.

Table 2: Back propagation parameter settings PLANNETT uses.

Parameter	Value
Input Units	one per input feature
Hidden Units	0, 5, 10, or 20
Output Units	2 (*volcano, nonvolcano*)
Activation Fn.	sigmoid
Training Epochs	500 (fixed)
Learning Rate	0.01
Momentum	0.90
Initial Weights	uniformly random in [-0.5, 0.5]

Experimental Results

These experiments compare the end-to-end sensitivity and specificity, with respect to the scientists' consensus labeling, of

- JARTOOL with current GAUSS classification module (JARTOOL-GAUSS)
- JARTOOL with new PLANNETT classification module (JARTOOL-PLANNETT)
- The two individual planetary geologists who created the consensus labeling (two trials each)

JARTOOL-GAUSS and JARTOOL-PLANNETT are identical except for their classification modules. The algorithms and the human scientists were all free to label any object in the four images as a volcano. For analysis, we present several graphs decomposing JARTOOL-PLANNETT's full set of 32 ANNs into smaller groups to investigate the impact of combining several input representations and numbers of hidden units.

The GAUSS and PLANNETT classifiers were evaluated via four-image cross validation. The FOA detections from each image in turn were held aside as an unseen testing set, while the classifier was trained on the FOA detections from the remaining three images, labeled according to the scientists' consensus. The algorithm performances shown are the aggregate results over the four testing sets.

Figures 1–5 (described in detail below) present the experimental comparison results. These graphs show the detection rate (sensitivity) versus false alarm rate (a measure of specificity) of JARTOOL-GAUSS, JARTOOL-PLANNETT, and smaller groups of ANNs as the threshold for classifying an object as a volcano is varied.[2] All graphs also include performance data for the two planetary geologists scored individually against the ground-truth consensus labeling in each of two trials. Because the same two scientists created the consensus labeling, the points plotting their individual performances may be optimistic.

[2] GAUSS's threshold applies to its computed posterior probability that an object is a volcano, and is qualitatively the same as PLANNETT's.

On each graph, the point (0, 100) represents perfect agreement with the consensus labeling, and the ideal curve is the line segment from (0, 0) to (0, 100) and the half line from (0, 100) to (∞, 100). The axes are scaled by the number of consensus volcanos, 163. Thus, 100% on the y axis indicates that all 163 consensus volcanos have been detected, and 100% on the x axis that 163 false alarms have been generated. The FOA module detected only 144 of the 163 consensus volcanos, so the algorithms' curves asymptote at $y = 88.3\%$ (144/163). There were also 481 false alarms (nonvolcanos) in the output of the FOA module, bring the total number of image patches GAUSS and PLANNETT classified to 625.

Figure 1 compares JARTOOL-PLANNETT, which combines all 32 ANNs (eight input representations, four numbers of hidden units each), to JARTOOL-GAUSS and the planetary geologists. Note the appreciable variation between the scientists and *within* each scientist's own labelings carried out at different times. The scientists' individual points are also quite distant from the consensus point (0, 100), despite the fact that the same two scientists created the consensus. This is thus a difficult problem even for knowledgeable experts. We therefore treat the individual scientist points as benchmarks of expert-level performance, and our goal is for our algorithms to achieve similar accuracy. As Figure 1 shows, JARTOOL-PLANNETT does achieve the same level of accuracy as Scientist 2 on this problem. This is the first algorithm to reach expert-level performance on this task, and represents a significant advancement of automating volcano cataloging. JARTOOL-PLANNETT improves upon the previous JARTOOL-GAUSS system, producing substantially more true detections for a given false alarm rate for most of the curve (see figure).

There is a sensitivity-specificity tradeoff as the volcano detection threshold is varied. For this work, we are most interested in the portions of the curves near the scientists' points as well as the overall picture of how the curves from two algorithms compare. If one assigns specific costs to missed volcanos and false alarms, cross validation techniques could be used to choose an appropriate threshold automatically.

Figure 2 breaks JARTOOL-PLANNETT's 32 ANNs into eight groups of four according to the input representation used to train. Each group combines four ANNs having 0, 5, 10, and 20 hidden units, respectively. The figure also shows JARTOOL-PLANNETT's curve combining all 32 ANNs. The different input representations produce a wide range of ANN accuracies. It is striking that JARTOOL-PLANNETT's unweighted averaging of all 32 ANNs yields *better* accuracy for much of the curve than any individual representation, despite including three very poor representations. This may indicate that the different representations succeed in producing ANNs whose errors are relatively uncorrelated or negatively correlated, a situation favorable to reducing error when combining models (Krogh &

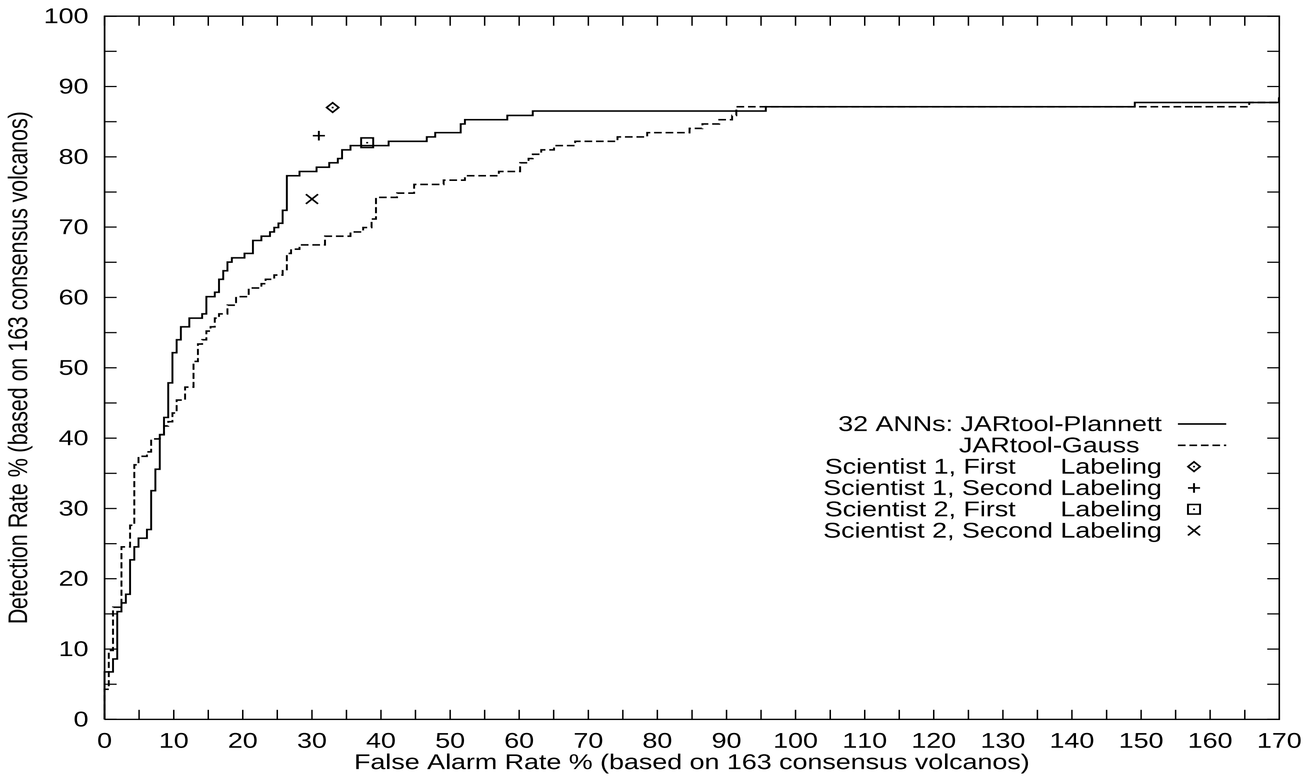

Figure 1: Comparison of JARtool-Plannett (32 combined ANNs) to JARtool-Gauss and human scientists.

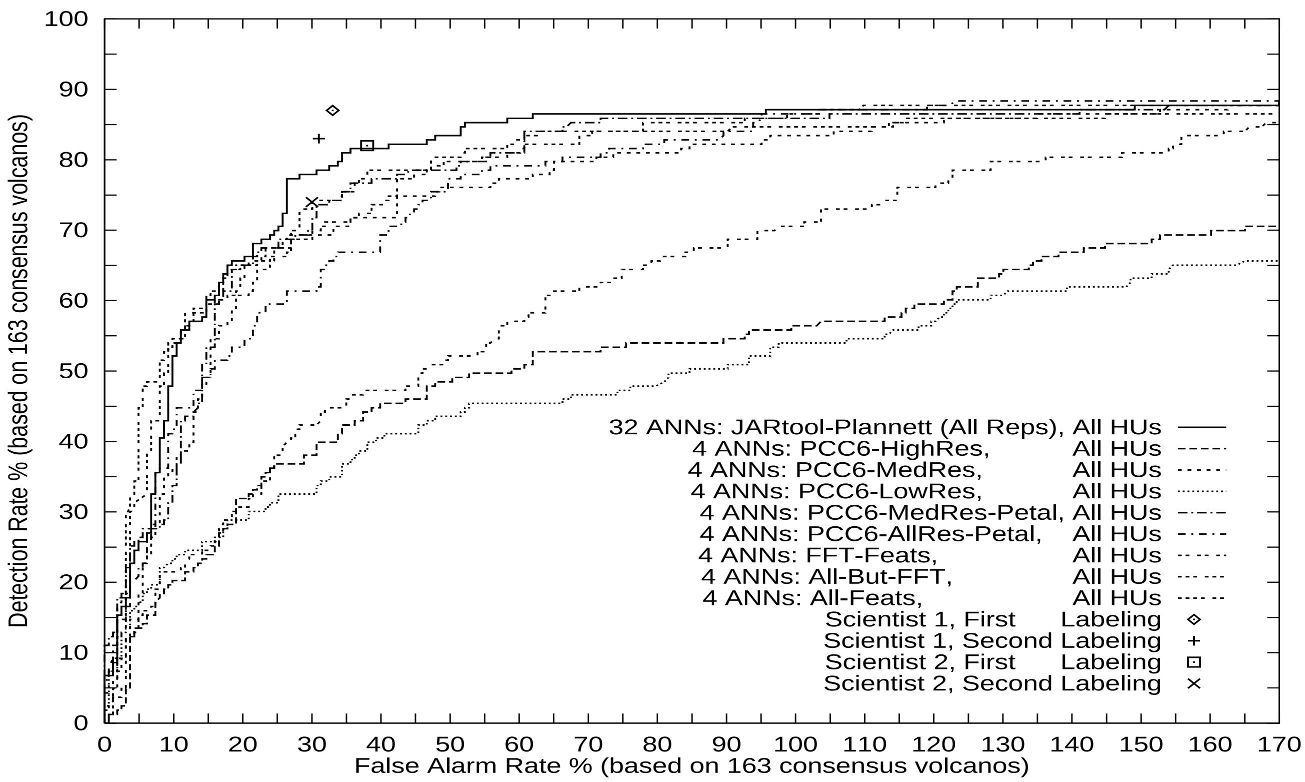

Figure 2: Comparison of ANNs trained on each individual input representation and combined (four ANNs per representation, with 0, 5, 10, and 20 hidden units, respectively) to JARtool-Plannett, which combines the ANNs from all eight representations (32 ANNs), and human scientists.

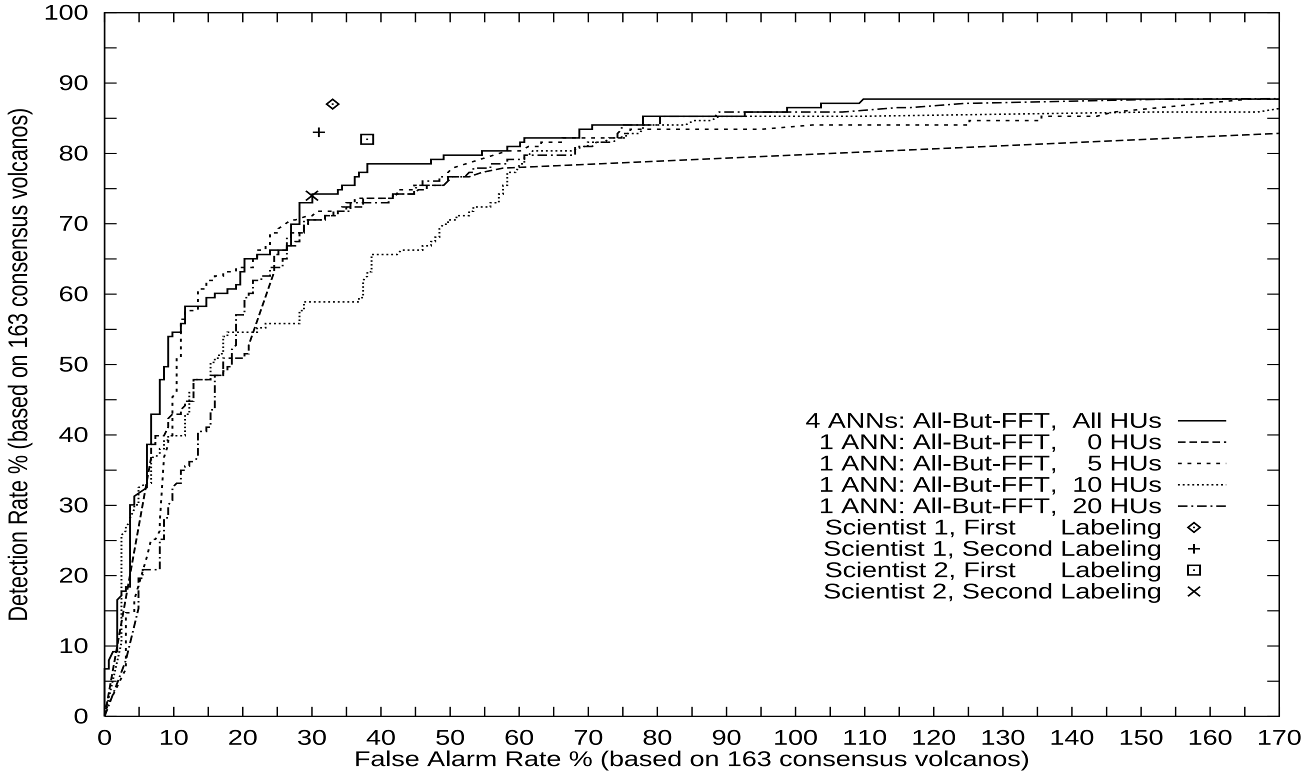

Figure 3: Comparison of combining four ANNs trained on representation All-But-FFT with 0, 5, 10, and 20 hidden units, respectively, to each individual ANN and human scientists.

Vedelsby 1995; Ali 1996). Interestingly, PCC6-MedRes is a very good representation, but the variations PCC6-LowRes and PCC6-HighRes, which use different source image resolutions, are the two worst representations. (The accuracy differences for these three representations are much smaller for the Gaussian classifier, but the ordering is the same.)

Figure 3 gives a more detailed picture of the ANNs trained on the All-But-FFT representation, which is the best individual representation as judged from Figure 2. It compares the curves for the four individual ANNs to the curve for the four ANNs combined (which also appeared in Figure 2). Individual ANN accuracy varies substantially for different numbers of hidden units, while the combined-ANN result here is actually better than that of the best individual ANN, which had 5 hidden units. We typically found that combining ANNs trained on a single input representation with different numbers of hidden units gave results about as good as the best individual ANN. This is fortunate, because the number of hidden units that produces the best results changes with different input representations. It would be unfair to look at the testing set results to choose the number of hidden units. Instead, by averaging over several numbers of hidden units, we do about as well as any individual network without having to choose a specific network size.

Figure 4 demonstrates that slight changes in the input representation can have a large impact on ANN classification accuracy. The figure shows curves for combining the four ANNs trained on each of the PCC6-MedRes-Petal and PCC6-MedRes input representations, respectively (two of the curves shown in Figure 2). Recall that PCC6-MedRes-Petal is the representation used by JARTOOL-GAUSS, containing six PCCs plus the petal feature. PCC6-MedRes is identical except for the omission of the petal feature. The figure shows that including the petal feature in this case substantially improves ANN accuracy.

Figure 5 compares the combined-ANN curve for representation PCC6-MedRes-Petal from Figure 4 to JARTOOL-GAUSS and the human scientists. This comparison is interesting because it varies the classification algorithm (ANNs versus Gaussian classifier) while holding the input representation, PCC6-MedRes-Petal, constant. For most of the curve, including the "knee" portion where the scientist points are, the combined ANNs slightly outperform JARTOOL-GAUSS. The single-ANN curves for this representation (not shown) cluster tightly around the two plotted curves. They are not shown simply because including them makes it nearly impossible to distinguish the intertwining JARTOOL-GAUSS curve. In this case, the combined ANNs perform better than the three individual ANNs having 5, 10, and 20 hidden units, and slightly worse than the ANN with no hidden units. On average, the individual ANNs using just the PCC6-MedRes-Petal representation are about as good as

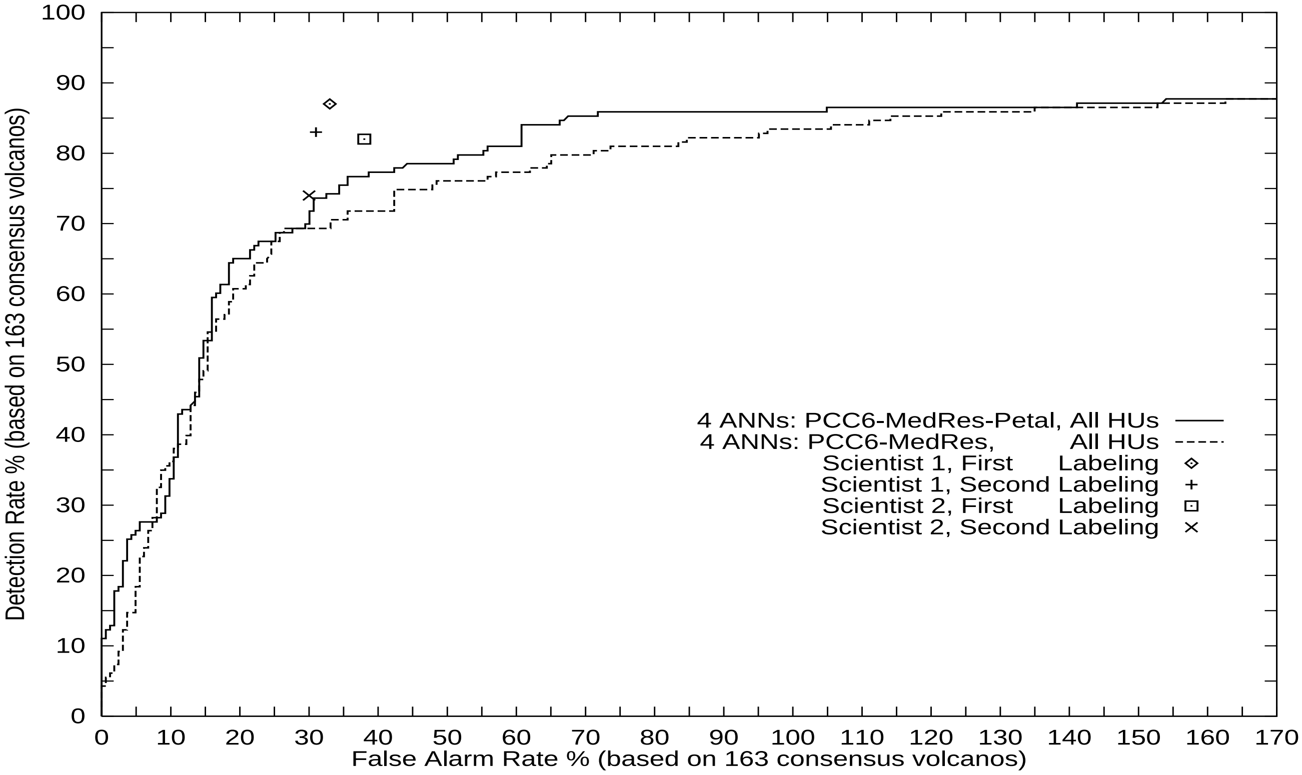

Figure 4: Comparison of combining four ANNs trained on representations PCC6-MedRes-Petal (seven features) versus PCC6-MedRes (six features) with 0, 5, 10, and 20 hidden units and human scientists. The representations differ from each other only as to whether or not the maximum petal response feature is included.

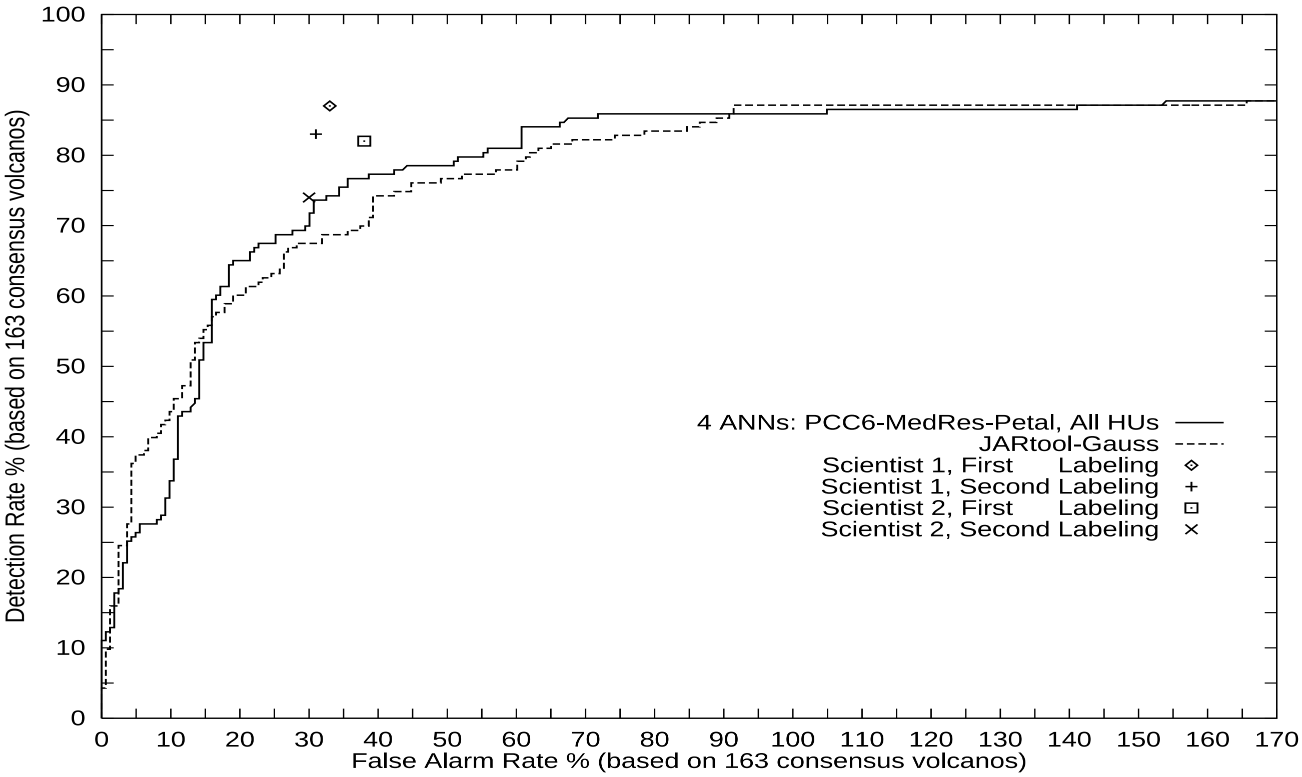

Figure 5: Comparison of combining four ANNs trained on representation PCC6-MedRes-Petal with 0, 5, 10, and 20 hidden units to JARtool-Gauss, which uses the same input representation, and human scientists.

JARtool-Gauss using the same representation (one is appreciably better, one is about equivalent, and two are slightly worse), while the combination of all four ANNs performs slightly better than average and has a small edge over JARtool-Gauss. This suggests that JARtool-Plannett's accuracy advantage over JARtool-Gauss is not primarily due to the different learning algorithm used (ANNs versus Gaussian classifier), but rather to its combination of many different input representations, although this hypothesis requires verification via experiments in which Gaussian models are trained and combined over the same eight input representations the ANNs used.

One advantage the Gaussian classifier has over ANNs is its much lower computational expense for small input representations. Gauss required less than one CPU second on a modern workstation to perform the entire four-image cross validation training and testing, while Plannett used about 2,700 seconds on comparable hardware to train and test the subset of its ANNs that use the same input representation as Gauss. For the entire experiment using all eight input representations, Plannett consumed approximately 45,100 seconds (about 12.5 hours). However, training time for the ANNs increases only linearly in the number of input features, while for the Gaussian classifier it increases approximately as the cube of the number of features, so Gauss's time advantage quickly disappears as larger feature sets are used. In any case, a few hours of extra computational expense are of little importance in a project that has spanned several years and in which final classification accuracy is the most important goal.

Conclusions

This paper introduced the Plannett system, which combines multiple ANNs to improve classification performance on the task of recognizing volcanos in Magellan radar images of Venus. Plannett replaces the current Gauss classification module of the JARtool image analysis system developed at the Jet Propulsion Laboratory. The resulting system, JARtool-Plannett, achieves the accuracy (both sensitivity and specificity) of a human planetary geologist at recognizing volcanos in a test suite of four images. JARtool-Plannett outperforms the JARtool-Gauss system and achieves the best algorithmic results on this image suite to date.

The analysis presented indicates that Plannett's accuracy improvements over Gauss likely stem from its combination of multiple different input representations and its use of more diverse sets of input features. Additional experiments are needed to determine whether the Gauss classifier could benefit equally from a similar combination scheme. Future work will test JARtool-Plannett and JARtool-Gauss on a larger set of labeled images that have not been used for algorithmic development. If results are favorable, Plannett may replace Gauss as JARtool's classification module of choice.

Acknowledgements

This research would not have been possible without the assistance of P. Smyth, L. Asker, U. Fayyad, V. Gor, M.P. Burl, M.C. Burl, J. Roden, and other current and former members of the Machine Learning Systems group at JPL. Special thanks to P. Smyth and L. Asker for providing the performance data for the Gauss classifier. This work was supported by a NASA Graduate Student Researchers Program fellowship.

References

Ali, K. 1996. *Learning Probabilistic Relational Concept Descriptions.* Ph.D. Dissertation, Information and Computer Science Dept., Univ. of California Irvine, Irvine, CA.

Burl, M.; Fayyad, U.; Perona, P.; Smyth, P.; and Burl, M. 1994. Automating the hunt for volcanoes on Venus. In *IEEE Computer Society Conf on Computer Vision and Pattern Recognition: Proc*, 302–309. Seattle, WA: IEEE Computer Society Press.

Clemen, R. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5:559–583.

Fayyad, U., and Smyth, P. 1993. Image database exploration: Progress and challenges. In *Proc Knowledge Discovery in Databases Wkshp, 11th Nat Conf on Artificial Intelligence.* Washington, DC: AAAI Press.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, 231–238. Denver, CO: MIT Press.

Pappas, T. 1992. An adaptive clustering algorithm for image segmentation. *IEEE Transactions on Signal Processing* 40(4):901–914.

Rumelhart, D.; Hinton, G.; and Williams, R. 1986. Learning internal representations by error propagation. In Rumelhart, D., and McClelland, J., eds., *Parallel Distributed Processing*, volume 1. Cambridge, MA: MIT Press. 318–363.

Smyth, P.; Fayyad, U.; Burl, M.; Perona, P.; and Baldi, P. 1995. Inferring ground truth from subjective labelling of Venus images. In *Advances in Neural Information Processing Systems*, volume 7, 1085–1092. Denver, CO: MIT Press.

Tresp, V., and Taniguchi, M. 1995. Combining estimators using non-constant weighting functions. In *Advances in Neural Information Processing Systems*, volume 7, 419–426. Denver, CO: MIT Press.

Wolpert, D. 1992. Stacked generalization. *Neural Networks* 5:241–259.