USING MACHINE LEARNING TO CLASSIFY IMAGE FEATURES FROM CANINE PELVIC RADIOGRAPHS: EVALUATION OF PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS AND ARTIFICIAL NEURAL NETWORK MODELS

FINTAN J. MCEVOY, JOSÉ M. AMIGO

As the number of images per study increases in the field of veterinary radiology, there is a growing need for computer-assisted diagnosis techniques. The purpose of this study was to evaluate two machine learning statistical models for automatically identifying image regions that contain the canine hip joint on ventrodorsal pelvis radiographs. A training set of images (120 of the hip and 80 from other regions) was used to train a linear partial least squares discriminant analysis (PLS-DA) model and a nonlinear artificial neural network (ANN) model to classify hip images. Performance of the models was assessed using a separate test image set (36 containing hips and 20 from other areas). Partial least squares discriminant analysis model achieved a classification error, sensitivity, and specificity of 6.7%, 100%, and 89%, respectively. The corresponding values for the ANN model were 8.9%, 86%, and 100%. Findings indicated that statistical classification of veterinary images is feasible and has the potential for grouping and classifying images or image features, especially when a large number of well-classified images are available for model training. © 2012 *Veterinary Radiology & Ultrasound.*

Key words: image classification, logistic regression, neural network, partial least squares discriminant analysis.

Introduction

MULTI-IMAGE STUDIES ARE increasingly being acquired in veterinary patients and computer-assisted diagnosis techniques may be helpful for increasing radiologist's efficiency and accuracy. One of the basic components of the diagnostic process is differentiation of one tissue from its neighbor. Statistical and computer software tools can be used to assist in this process and many such tools are based on the concepts of machine learning. The term machine learning is used when a computer program is applied to a well-posed problem and has a measurable performance that improves with experience.[1] Machine learning algorithms are commonly used in applications such as face recognition software,[2] internet search engines,[3] and spam filters[4] to detect patterns in data and perform classification tasks. Machine learning algorithms have been used for a variety of geology, astronomy, and medical imaging applications.[5–8] In veterinary medicine, machine learning algorithms have been used for applications such as lameness diagnosis in cattle,[9] epidemiology,[10] and food processing.[11] We hypothesized that these algorithms may also be used for veterinary diagnostic imaging applications.

Partial least squares discriminant analysis (PLS-DA) models are machine learning classifier systems based on logistic regression. They have been used for image classification problems such as detecting lesions consistent with Alzheimer's disease in human brain magnetic resonance imaging studies.[12] The method generates a linear model and classification boundaries. Artificial neural network (ANN) models are also logistic regression classifiers that have been used for image classification problems such as detecting lesions consistent with neoplastic nodules in human breast ultrasound studies.[13] An advantage of ANNs is that they are able to generate nonlinear models. Partial least squares discriminant analysis and ANN models both require training and cross-validation (test) image sets. Training sets are used to iteratively optimize the model parameters and test image sets are used to measure classification ability (accuracy) of the final model.[1]

The purpose of this study was to evaluate linear PLS-DA and nonlinear ANN models as classifier systems for identifying regions of interest containing the hip joint on ventrodorsal pelvis radiographs in dogs.

From the Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences (McEvoy) and Department of Food Science, Quality and Technology, Faculty of Science (Amigo), University of Copenhagen, Copenhagen, Denmark.

Part of this study was presented in abstract form at the 16th International Veterinary Radiology Association meeting, Bursa, Turkey, 2012.

Funding source: Chemometric Analysis Center (CHANCE), University of Copenhagen, Copenhagen, Denmark.

Address correspondence and reprint requests to Fintan J. McEvoy, at the above address. E-mail: fme@sund.ku.dk

Received May 11, 2012; accepted for publication October 3, 2012.
doi: 10.1111/vru.12003

FIG. 1. A random selection of nine images from the training set. Examples of images labeled as hip, that is assigned a value "1," are seen in the top row, left image; middle row, left and right images; bottom row, left image. The image in the middle row, middle position, contains part of the hip, but this was labeled in the training set as not a hip, value "0," as the image does not meet the criteria set for a hip image.

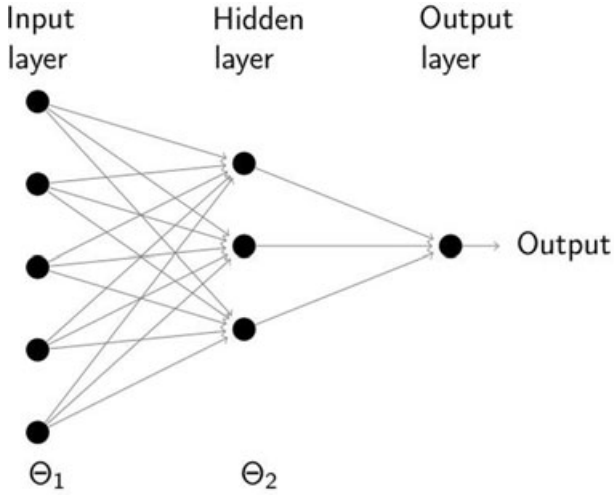

FIG. 2. An artificial neural network, comprising five nodes in the input layer, three nodes in a single hidden layer, and a single node in the output layer. Θ_1 and Θ_2 are parameters of the model; their values are randomly set at the outset and modified iteratively to improve accuracy. In the network illustrated, Θ_1 will be a 5×3 matrix and Θ_2 a 3×1 matrix. The network used in this study had three layers and a single output node as illustrated, but unlike the illustration had 4096 input nodes and 25 hidden layer nodes.

Materials and Methods

Training and test sets of images were created using consecutive ventrodorsal pelvis radiographs of 60 dogs. Images were retrieved from the Picture Archiving and Computer System for the University of Copenhagen's Veterinary Teaching Hospital. For each side of the pelvis, a region of interest (ROI) was created to include the tuber ischium, full extent of the hip joint, and lateral edge of the obturator foramen. The cranial and lateral margins of these ROIs were set so that the final ROI was square. The ROIs were saved as separate 64×64, 8-bit, portable network graphics (.png) format, gray scale images. These ROIs were defined as hip or positive outcome. Nonhip or negative outcome ROIs of identical dimension were made from randomly chosen areas on the ventrodorsal radiographs that did not include the hip as described above. These ROI images contained other structures such as abdominal soft tissues, vertebrae, the stifle or part, but not all, of the hip. Images of hip and nonhip areas were right-left transposed in 18 of the dogs and added, without change of classification, to the image sets. The resulting data sets consisted of 200 training images (120 hip and 80 nonhip) and 56 test images (36 hip and 20 nonhip). Examples from the data set are shown in Fig. 1.

Partial least squares discriminant analysis was performed using PLS-Toolbox v3.5 (Eigenvector Research, Wenatchee, WA) implemented in MATLAB v7 software (Mathworks, Natick, MA). The model was trained with a matrix containing the 200 previously described training images. Each 64×64 pixel image was converted to a 1×4096 vector so that the input data comprised a 200×4096 matrix. The corresponding classifications for these images were input as a 1×200 matrix, containing values "1" for hip regions, and "0" for nonhip regions.

The general structure of the ANN model comprised an input layer with 4096 logistic units or nodes, a single hidden layer with 25 nodes, and an output layer containing one node (i.e., the output of the model) (Fig. 2). The 64×64 matrix images were again converted to a 1×4096 vector. The vector was inserted into the input layer, with one vector element per input node. A sigmoid (logistic regression) function was used in determining activations of the hidden and output layers. The activation parameter Θ_1, from input layer to hidden layer, and the parameter Θ_2, from hidden layer to output layer, were randomly chosen to initialize the model and optimized in 200 iterations using an optimization function.[14] The programming code required to implement the network was written in "Octave" (GNU Octave http://www.octave.org). In this way, the network was optimized on the training set. The network was designed to output a probability value between 0 and 1 that a given input (image) contained a hip. If this probability value was greater than or equal to 0.5, the image was classified as "hip," otherwise the classification was "nonhip."

The set of 56 test images was used to test the models and to calculate measures of performance. The error for each

Table 1. Performance of a Partial Least Squares Discriminant Analysis (PLS-DA) Model and an Artificial Neural Network (ANN) Model as Image Classifiers*

	PLS-DA		ANN	
	Training	Test	Training	Test
Sensitivity	0.925	1.000	0.992	0.86
Specificity	0.983	0.889	0.875	1.000
Classification error	0.046	0.067	0.055	0.089

*Performance indices for the training set (200 images) and the validation (test) set (56 images) for each model.

classification was determined by subtracting the probability output by the model from the actual value (1 for a hip image and 0 for nonhip). The mean of the squared error for all test images was then determined. Plots were created for mean squared error vs. latent variable number for the PLS-DA model and for mean squared error vs. the number of iterations for the neural network model. Classification error, specificity, and sensitivity[15] for the models generated by PLS-DA and the ANN models were determined for both the training and test images sets.

Results

Partial least squares discriminant analysis model and ANN model comparisons are summarized in Table 1. Partial least squares discriminant analysis training using three latent variables produced acceptable classification performance. This number was determined as optimal from a plot of the classification error vs. the number of latent variables (Fig. 3). The classification error, specificity, and sensitivity on the test set were 6.7%, 89%, and 100%, respectively.

Training the neural network using random initializations (random Θ_1 and Θ_2) yielded training set classification errors of between 15% and 4%. Processing time for training was between 2 and 4 min depending on the computer used (PC, Intel processor (dual core), 8 GBytes ram, Linux (Fedora) and MacBook Pro, PowerPC processor, 2 GBytes ram, MacOS X. Faster processing times on the former.). Optimal model parameters yielded a classification error of 8.9%, sensitivity of 86%, and specificity of 100% for the test set. Using the pretrained activation parameters, image classification was achieved in an average time of 2.19 ms per image. Learning curves demonstrating mean squared error vs. number of iterations are shown in Fig. 4. With increasing number of iterations there was a better fit between the model and the training set, in the range tested (20–500 iterations). In the test set, however, iteration numbers in the region 200 gave optimal performance, with no clear improvement available from further iterations.

Discussion

Findings from our study indicated that common machine learning algorithms can be used to classify veterinary radiographic images. The classification process was so powerful that, for both models, no a priori image feature identification, so-called use of prior knowledge, was needed for a successful implementation. Only the pixel data and the training set classifications were required. Partial least squares discriminant analysis is a classification technique that builds a linear model to predict outcome using all the training images arranged in a matrix. This matrix is the sample or X matrix. A second matrix is constructed containing the classifications for the training set. In this study, the classifications were "hip" and "nonhip" and the values in the matric were either 1 or 0, respectively. This second matrix was an outcome or Y matrix. The X matrix had many variables (one for each pixel value in each training image). The model iteratively created new smaller X matrices that contained fewer variables (called latent variables) than the original but yet retained as much as possible of the variation that distinguishes between images of different classes. The model then found linear boundaries that could be used to separate the classes (1 or 0) based on the probability of an image belonging to a particular class. Once the model was trained, the classification boundaries were set and could be used to classify test images. The size of the training set is critical and it is usual that the greater the number of training examples the better the performance. A balance has to be found, however, between the improved performance and the cost of obtaining more samples. Parameters describing image features, rather than the raw image data itself can be used as replacement or additional inputs to a neural network. This approach has been used to classify lung nodules as benign or malignant and resulted in a classification accuracy of 90.3%.[16]

In this study, original images and right-left transpositions of these were used as input. This is common practice in machine learning implementations, where there is an expectation for the program to recognize objects irrespective of orientation.[17] We did not include up/down transpositions. Our rational was that the model should be able to recognize both the right and left hip on radiographs presented in normal viewing orientation, but not necessarily on images presented upside down. Partial least squares discriminant analysis was able to classify the test samples with high values of sensitivity and specificity and low error rate. This finding indicated that, despite the few samples available compared with the number of variables; a linear model could classify hip images in the scenario used. The ANN model likewise performed well. While ANN models are capable of generating complex nonlinear regression models, they have a tendency to over fit data with the result that high accuracies are achieved in training sets yet

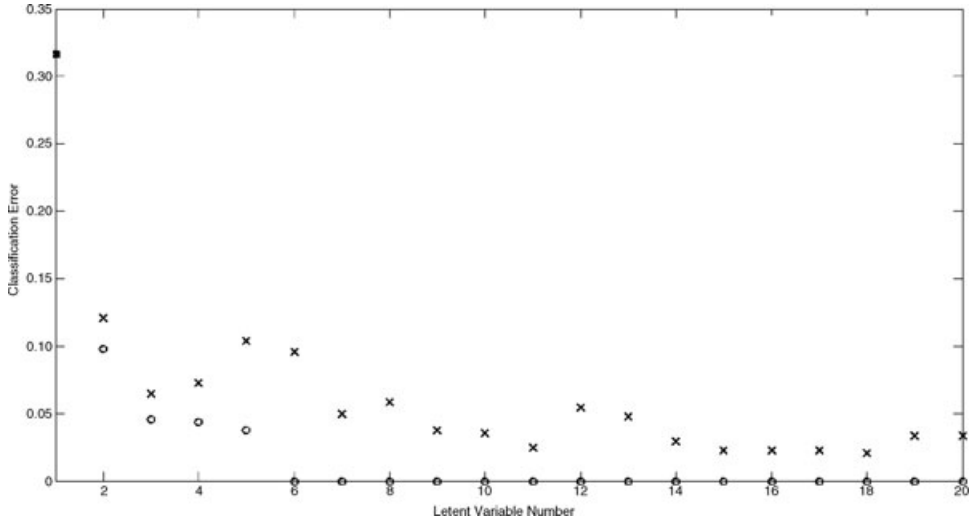

FIG. 3. Learning curve, classification error for the training set (circles), and test set (crosses) against the number of latent variables for the partial least squares discriminant model. The model error on the test set initially decreases for the first three latent variables. Thereafter, the effect of increasing latent variable number on model performance levels off.

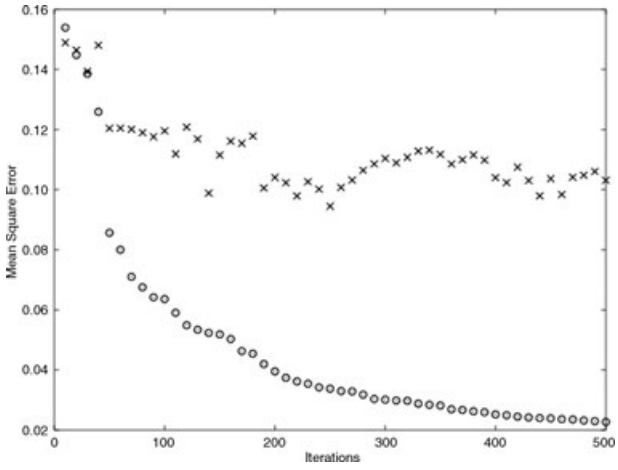

FIG. 4. Learning curve, classification error against the number of iterations for the artificial neural network model. The training set error (circles) decreases toward zero with increasing iterations. Test set error (crosses) decreases with iteration number, but this effect levels off after 200 iterations.

performance can be poor for previously unseen, test images. This is especially so with large neural networks, containing many parameters, such as used in this study. The good performance results in this implementation, however, suggest that overfitting was not a problem. Other aspects of ANN architecture, for example, the transfer function applied to the node activations, can be altered to improve performance if needed. A sigmoid function was used in this study but others are available. Likewise the number of hidden layers can be altered, one or more can be used, and also the number of nodes in each hidden layer must be determined. The choices made in regard to these ANN features affect the network's performance.

A particular weakness in the implementation of the ANN in this study was the sample per feature ratio. With only 200 samples and 4096 features, the ratio (0.05) was far removed from the ideal (5–20). The problem could manifest as a lack of robustness in that the performance of the model could not be guaranteed as more test cases became available. For future studies, the depth of understanding required to modulate networks in these and other ways will require collaboration, such as used in this study, between clinical imaging and model specialists.

The classification task in this study was binary, with input being classified as a hip or not. The models generated could be used to find ROIs containing hips on radiographs, a process analogous to face detection on a photograph. For future studies, PLS-DA and ANNs could also be used as multiclass classifiers. In a multiclass classifier design, all inputs could be images of hips, perhaps selected automatically by one of the models developed here. Classification into one of 6 classes could then be modeled (class 1 as normal; classes 2–6, increasing degrees of abnormality), perhaps using an ANN. In medical imaging, machine learning has been previously used with some success as a diagnostic aid in breast cancer using ultrasound[12] and mammography[18] images and in lung nodule detection on computer tomography images.[19] In these and other medical applications, the role of the algorithm was to assist diagnosis. There is a trend in medical imaging toward increased image data set sizes, e.g. progression from single to multirow detector computer tomography or from 2D to 3D ultrasound. This trend and the large digital libraries associated with modern medical imaging make computer aids particularly attractive. In practice, these may be used to suggest a

particular classification to retrieve images similar to a given input image or to group or cluster similar images. Despite their impressive classification powers, machine learning applications in medical and veterinary imaging are not common. A lack of sufficiently large numbers of well-classified images, difficulties in accessing images, particularly in human medical practice, a lack of the skills among radiologists that are necessary to get a foothold in the area and poor communication between experts in the field of machine learning and radiologists, are all possible contributing factors.

In conclusion, findings from this study indicated that machine learning techniques can be applied to veterinary radiographic images. Future studies are needed to further refine these techniques and determine their utility for computer-assisted diagnosis, particularly in research settings where large numbers of well-classified digital images are available.

ACKNOWLEDGMENTS

The authors are grateful to the Chemometric Analysis Center (CHANCE), University of Copenhagen, for financial support.

REFERENCES

1. Mitchell TM. Machine learning. Boston, USA: McGraw-Hill, 1997.

2. Ginhac D, Yang F, Liu X, Dang J, Paindavoine M. Robust face recognition system based on a multi-views face database. In: Grgic M, Delac K, Bartlett MS (eds): Recent advances in face recognition. New York, USA: InTech Publishers, 2008;37–38.

3. Cordón O, Herrera-Viedma E, Lápez-Pujalte C, Luque M, Zarco C. A review on the application of evolutionary computation to information retrieval. Int J Approx Reason 2003;34:241–264.

4. Guzella TS, Caminhas WM. A review of machine learning approaches to spam filtering. Expert Syst Appl 2009;36:10206–10222.

5. Mjolsness E, DeCoste D. Machine learning for science: state of the art and future prospects. Science 2001;293:2051–2055.

6. Hoi SC, Jin R, Zhu J, Lyu MR. Batch mode active learning and its application to medical image classification. *Proceedings of the 23rd International Conference on Machine learning, ICML '06*, Pittsburgh, Pennsylvania, USA, 2006;417–424.

7. Bhooshan N, Giger M, Edwards D, et al. Computerized three-class classification of MRI-based prognostic markers for breast cancer. Phys Med Biol 2011;56:5995–6008.

8. Suzuki K. Pixel-based machine learning in medical imaging. Int J Biomed Imaging 2012;2012:792079.

9. Ghotoorlar SM, Ghamsari SM, Nowrouzian I, Ghotoorlar SM, Ghidary SS. Lameness scoring system for dairy cows using force plates and artificial intelligence. Vet Rec 2012;170:126.

10. Garcia Alvarez L, Webb C, Holmes MA. A novel field-based approach to validate the use of network models for disease spread between dairy herds. Epidemiol Infect 2011;139:1863–1874.

11. Iqbal A, Valous N, Sun DW, Allen P. Parsimonious classification of binary acunarity data computed from food surface images using kernel principal component analysis and artificial neural networks. Meat Sci 2011;87:107–114.

12. Westman E, Simmons A, Muehlboeck JS, et al. AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. Neuroimage 2011;58:818–828.

13. Joo S, Yim HC. Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features. IEEE Trans Med Imaging 2004;23:1292–1300.

14. Rasmussen. CE. Available at http://sprinkler.googlecode.com/svn/trunk/regression/fmincg.m, 2012.

15. Pérez-Bernal JL, Amigo Rubio JM, Fernández-Torres R, Bello M, Callejón-Mochón M. Trace-metal distribution of cigarette ashes as marker of tobacco brands. Forensic Sci Int 2011;204:119–125.

16. McNitt_Gray MF, Hart EM, Wyckoff N, Sayre JW, Goldin JG, Aberle DR. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. Med Phys 1999;26:880–888.

17. Stojmenovic M. Real time machine learning based car detection in images with fast training. Mach Vision Appl 2006;17:163–172.

18. Ge J, Sahiner B, Hadjiiski LM, et al. Computer aided detection of clusters of microcalcifications on full field digital mammograms. Med Phys 2006;33:2975–2988.

19. Tan M, Deklerck R, Jansen B, Bister M, Cornelis J. A novel computer-aided lung nodule detection system for CT images. Med Phys 2011;38:5630–5645.