A Comparison of Backpropagation and Statistical Classifiers for Bird Identification

A.L. McIlraith and H.C. Card
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba, Canada R3T 5V6
hcard@ee.umanitoba.ca

Abstract

We compare neural networks and statistical methods used to identify birds by their songs. Six birds native to Manitoba were chosen which exhibited overlapping characteristics in terms of frequency content, song components and length of songs. Songs from multiple individuals in each species were employed. These songs were analyzed using backpropagation learning in two-layer perceptrons, as well as methods from multivariate statistics including quadratic discriminant analysis. Preprocessing methods included linear predictive coding and windowed Fourier transforms. Generalization performance ranged from 82% to 93% correct identification, with the lower figures corresponding to smaller networks that employed more preprocessing for dimensionality reduction. Computational requirements were significantly reduced in the later case.

1. Introduction

Recently, interest has arisen in the possibility of automatic species identification of bird sounds by computer. An advantage of this approach is that it would not require an expert user. Such technology could be employed for long-term environmental monitoring [1] and education. The task of automatic species recognition is difficult because bird songs are variable, as becomes evident when one tries to learn to identify birds by their sounds.

Analysis of bird sounds usually begins with the creation of frequency-time representations of recordings [2,3]. Temporal and spectral features [4] are then obtained from this representation. Some traditional feature measurements have included the frequency having the highest energy content within an element [5], the number of elements per song and durations of elements and inter-element intervals [6]; we call a burst of sound within a song an element. Although there are many transformations for converting a time-domain signal into a frequency-time representation, we employ the windowed Fourier analysis [7] due to its wide use and conceptually simplicity. To obtain statistically valid spectral estimates using the FFT, power spectral densities [8] can be used. Another effective preprocessing method is linear predictive coding or LPC analysis, which was originally inspired by the source-filter model applied in speech analysis [9].

Once a set of feature variables is obtained for a set of songs, data can be presented to a classifier, or statistical methods such as stepwise discriminant analysis used to reduce dimensionality prior to classification. Classification can be performed using artificial neural networks and a variety of multivariate statistical methods [10]. When appropriately applied, both may be effective for a given problem [10-12]. We utilize and compare both types of classifiers.

2. Methods

133 songs from different individuals were extracted from audio tapes and compact disks [13-18]. Recordings varied in quality. Six bird species were chosen: Song Sparrow (SON), Fox Sparrow (FOX), Marsh Wren (MWR), Sedge Wren (SWR), Yellow Warbler (YLW) and Red-winged Blackbird (RWB). Songs were digitized with 8 bit resolution at 11.025kHz, with levels being adjusted to give maximum amplitude without clipping. Preprocessing and classification were performed in two ways, which we term methods one and two.

In method one, songs were framed using a non-overlapping Hamming window with 512 samples (46msec), a width similar to that used in speech recognition [19]. The 15th order LPC filter was

0-7803-4122-8/97 $10.00©1997 IEEE

determined for each frame and the FFT of the whitening filter coefficients used to generate nine unique spectral magnitudes. This procedure was repeated using 2048 sample windows. Nine spectral variables and a temporal variable (song length) were normalized to a mean of zero and a standard deviation of one. Data were compressed using a logistic function with unity gain. Two separate ten variable datasets were constructed for 512 and 2048 sample window sizes.

Backpropagation without momentum or higher order derivative information was employed as the learning model [20-21]. Experimentation suggested that a network of 10 inputs, 12 hidden nodes and six outputs was appropriate. The learning rate was 0.2 and target values were changed from 0.0 and 1.0 to 0.2 and 0.8 to accelerate learning [22]. Songs were divided between test and training data sets randomly. The proportion of test data was 25%. To evaluate performance, 10 data sets were generated for each window size, and the network trained with new initial weights each time. The 10 test and 10 training sets were generated in the same order for both window sizes. Preliminary training runs of 10000 epochs indicated that 1500 epochs was sufficient for the mean sum-of-squares error (MSE) to converge. For each test song, output activations exceeding 0.6 for its windows were totaled for the 6 species outputs. The species class with the largest count was considered to be the winning class. If two classes were tied for the maximum or if no class won, classification was considered to be incorrect.

In method two, the same songs were used. Temporal processing of sounds was performed using a 'leaky-integrator' algorithm that parsed songs in a manner that matched human perception of elements. As a result of adjustment, silences that were too short in duration to be perceptible were ignored. The number of elements in a song was determine and the mean and variance of both element and silence lengths calculated; means and standard deviations for measured quantities, and the element count constituted the five temporal variables extracted from each song. The amplitude of each element was normalized prior to spectral analysis. Power spectral densities were calculated for each element using the Welch method [8], a triangular window and a 16-point FFT. Magnitudes for nine spectral bands were averaged within each element. Band averages were normalized with respect to the band containing the largest average to reduce the effect of amplitude variations. Means and standard deviations for the nine spectral bands generated eighteen spectral variables for each song. SAS software was used for all subsequent statistical analyses [23,24] of the 23 variable data set.

Preliminary examination of the data correlation structure indicated complex inter-correlation of variables. This in turn suggested that a smaller set of variables could contain enough information for discrimination. Toward this end, stepwise discriminant analysis was performed; the significance criterion used to enter a variable and to retain it during the elimination phase was $\alpha = 0.15$. Results indicated that the number of variables required could be reduced from 23 to eight. The amount of overlap among species' songs was examined by reducing the eight variable data set to a two-dimensional space with canonical discriminant analysis (CDA).

Discriminant analysis was used to classify songs. Quadratic discrimination analysis (QDA) was necessary since covariance matrices for different classes were too dissimilar to pool. The records were divided in a stratified random manner into a test set containing one record of each species, and a training set which contained the remaining 127 records. Training data were used to generate discriminant functions, which were then used to classify the 'unknown' test songs. In order to reduce the impact of atypical records, 25 different randomizations were used. Using the same data, the backpropagation learning rule was applied with an initial learning rate of 0.2, and target values of 0.1 and 0.9. Eight inputs and six outputs were used. To determine the best network configuration, the number of hidden nodes was varied from three to eight, and the number of training epochs from 20 to 500. The best accuracy was obtained with 6 hidden units and 200 epochs. This configuration was used with 25 randomizations, as described above, to generate the final results. In testing, any for a given song with an output in excess of 0.6 was considered to be active. If any incorrect output was active or if no outputs were active, a misidentification was recorded. To make the results comparable to those derived from classification by discriminant analysis, training and test sets were generated in the same manner for both.

3. Results

The mean squared error (MSE) at the end of training was significantly larger for the 512 sample window data ($p \ll 0.0005$, n=10, with a two-tailed t-test [25]). The network was able to recognize the six species by their song, as shown in Fig. 1. Overall performance ranged from 91% to 93% correct identification. Except for FOX and RWB songs, mean performance was somewhat higher for the 2048 data sets than for the 512 data sets. Examination of raw data indicated that certain songs were consistently misidentified. In two cases, a song was either atypically long or short, and was misidentified regardless of the data set in which it was included.

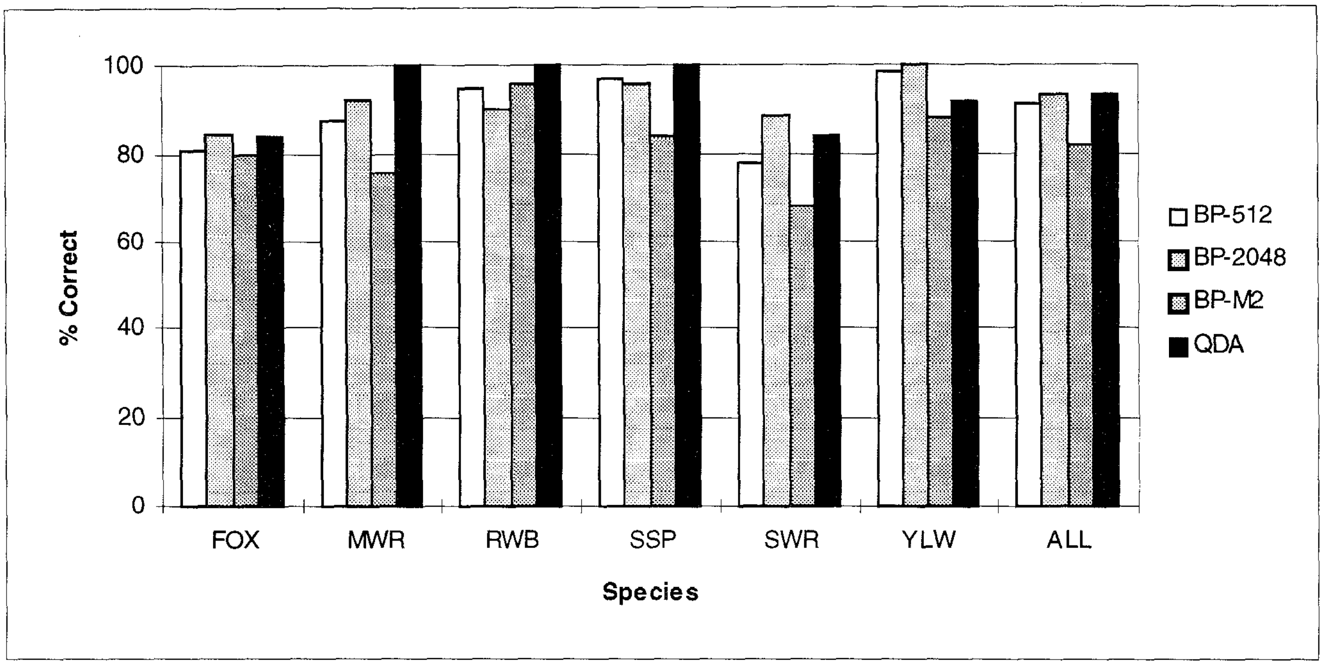

Figure 1. Species recognition performance for multi-layer perceptrons (with backpropagation learning) with 512 and 2048 sample window data of method one, for method two and for QDA. The category 'ALL' represents performance over all species.

Stepwise discriminant analysis suggested that the number of variables could be reduced from 23 to eight in subsequent analyses since the average squared canonical correlation increased only slightly (0.01 between steps eight and nine) if more than eight variables were retained. Those variables with large partial correlations (Table 1) and F-statistics contributed the most to discrimination. Note that all F-statistics were significant ($p \leq 0.0001$). XE and NS were the dominant temporal variables; S4 was the dominant spectral variable.

Table 1. Contribution of variables to the canonical discriminant model including the partial correlation coefficient for the variable and the corresponding F-value. CD1 and CD2 are the eigenvectors for the first two CDA axes. The two largest eigenvector components are indicated in bold type.

Variable	Partial r^2	F	CD1	CD2
NE	0.57	32.2	-.64	**0.58**
S4	0.59	35.0	0.60	0.41
SE	0.26	8.5	**0.92**	0.12
X2	0.39	15.4	0.73	0.48
X3	0.39	15.1	0.69	0.003
X8	0.26	8.4	-.66	**-.61**
XE	0.52	26.2	**0.94**	-.007
XS	0.42	17.1	-.63	0.56

Canonical discriminant analysis indicated strong data structure, accounting for 83% of standardized variance with only two axes, and all of it with five. Element length and standard deviation contributed the most to CD1 (Table 1). CD2 was affected by both temporal and spectral variables. YLW and RWB songs formed distinct groups (Fig. 2). Other species had varying degrees of overlap, with FOX and MWR songs being the least distinct.

QDA yielded an accuracy of 93.3% for test records (Fig. 1), and showed a similar pattern of variation to those observed for method one, while the 8-6-6 multi-layer perceptron had an accuracy of 82% (Fig. 1).

4. Discussion

The application presented in this study is novel, since we know of only one other published description of an artificial neural network designed to recognize bird calls [26]. Since the goal was to recognize classes of sounds from several individuals, the level of difficulty was similar to that of a speaker-independent word recognition task, with songs being recorded from individuals in a variety of locations, with differing quality and amounts of background noise. Given that this remains a challenging problem in human speech research, the overall performance achieved with both methods was good.

In method one, increasing the window size from 512 to

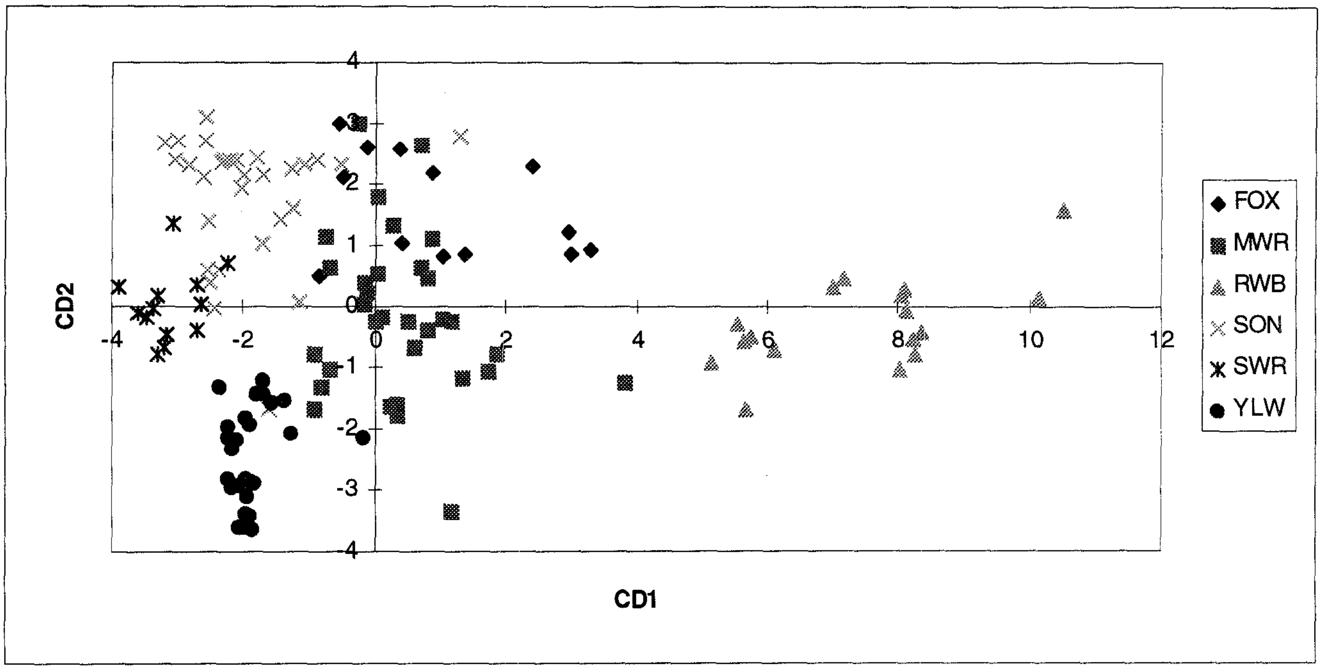

Figure 2. The first two canonical discriminants from the 8 variable data set, plotted by species.

2048 samples resulted in a slight improvement in neural network performance. This change added some implicit temporal context beyond that provided by the song length variable. Although method one provided excellent results, considerable computation was required. Training consumed several hours on a high-performance workstation due to the dimensionality and the number of input records generated in preprocessing. We doubt that this method would scale well with larger sample sizes and the addition of more species. For these reasons, method two was investigated. In method two, we generated a consistent and small number of variables for each song which described its temporal and spectral characteristics. Calculating the mean for measured parameter variables over all elements of a song summarized the central tendency of that parameter for the song. Standard deviations provided information about the parameter's variability within the song. Further dimensionality reduction was accomplished with stepwise discriminant analysis, in order to reduce the degrees of freedom required within the final classifiers.

The graphical representation of the multivariate data provided by CDA coupled with the large amount of variance captured in two axes suggested that there was a strong underlying structure in the data set, and that linear transformation of the 8 variables into 2 or 3 retained much of the information. CDA also indicated that the 8 variable data set contained considerable multivariate overlap, suggesting that this recognition task was tractable but not trivial.

While the results obtained with QDA were excellent. The multi-layer perceptron did not classify quite as well as QDA in method two, or as well as it did in method one. This we believe to be due to suboptimal preprocessing, rather than to an undersized network for classification. Had a larger unsupervised network been allowed to discover the best features from the underlying data, the multi-layer perceptron may have performed better. Of course, this would have greatly increased the overall computation and learning time as in method one. It is also possible that pathological cases were more prevalent among the test samples used in testing this network or that the method of judging the winning output was too conservative. In spite of the reduced performance, the smaller number of hidden units and shorter training periods in method two suggest that this approach has a place in applications.

5. Conclusions

Explicit temporal preprocessing and statistical methods were effective in reducing data dimensionality, and in turn permitted successful application of both statistical and neural classifiers. Statistics and neural networks are complementary tools for pattern recognition and classification research. Methods using multi-layer perceptrons can trade performance for computational efficiency, enabling various applications on desktop workstations or simple custom field hardware. In future work we will examine datasets with larger numbers of species and samples per species. We anticipate that as the recognition task becomes more difficult the relative contribution of the parameters to the discrimination model will change, with an overall performance decline. In this case we plan to use a hierarchical network of adaptive experts [27], drawing upon artificial neural network and statistical methods. We anticipate improving the neural network classifier using statistical methods as benchmarks, and using unsupervised neural learning methods to better automate preprocessing.

6. Acknowledgments

We wish to thank members of the Ganglion research group for their helpful input. This research was supported by NSERC and by a University of Manitoba Fellowship to A.L. McIlraith.

7. References

1. M. Mittelstaedt, "Ontario Hydro jolting employees into more productivity", *The (Toronto) Globe and Mail*, January 10, pp. A4, 1994.
2. T. Aubin, "Some features of time-frequency analysis and representation of animal vocalizations", in "The conference report of the XII Symposium of the International Bioacoustics Council", *Bioacoustics*, 4:59-60, 1992.
3. P.K. McGregor, "LSI's speech workstation: a sound analysis package for IBM PCs", *Bioacoustics*, 3:223-234, 1991.
4. N.S. Thompson, K. LeDoux and K. Moody, "A system for describing bird song units", *Bioacoustics*, 5:267-279, 1994.
5. D.M. Weary, R.G. Weisman, R.E. Lemon, T. Chin and J. Mongrain, "Use of relative frequency of notes by Veeries in song recognition and production", *Auk*, 108:977-981, 1991.
6. T. Rich, "Microgeographic variation in the song of the Sage Sparrow", *Condor*, 83:113-119, 1981.
7. P. Kraniauskas, "A plain man's guide to the FFT", *IEEE Sig. Proc. magazine*, 11:24-35, 1994.
8. W.H. Press, W.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge University Press, N.Y., 1992.
9. B.S. Atal, "Linear Predictive Coding of Speech", in *Computer Speech Processing*, F. Fallside and W.A. Woods, eds., Prentice-Hall, Inc., London, pp. 81-124, 1985.
10. L. Breiman, "Comment", added to "Neural networks: a review from a statistical perspective" by B. Cheng and D.M. Titterington, *Statist. Sci.*, 9:2-54, 1994.
11. B.D. Ripley, "Neural networks and related methods for classification", *J. R. Statist. Soc. Series B.*, 56:409-456, 1994.
12. B. Cheng and D.M. Titterington, "Neural networks: A review from a statistical perspective", *Statist. Sci.*, 9:2-54, 1994.
13. M. Brigham, *Bird Sounds of Canada*. Holborne Dist. Co. Ltd., Mount Albert, Ontario, no date.
14. D.J. Borror, *Songs of Eastern Birds*, Dover, N.Y., 1970.
15. D.J. Borror, *Common Bird Songs*, Dover, N.Y., 1967.
16. L. Elliot and T. Mack, *Wild Sounds of the Northwoods*, NatureSound Studio, Ithaca, N.Y., 1990.
17. R.K. Walton and R.W. Lawson, *Birding by Ear (Eastern / Central) - a Guide to Bird-Song Identification*, Houghton - Mifflin, Boston, 1989.
18. P.P. Kellogg, R.T. Peterson and W.W.H. Gunn, *A Field Guide to Western Bird Songs*, Houghton - Mifflin, Boston, 1975.
19. R.P. Lippmann, "Review of neural networks for speech recognition", *Neural Computation*, 1:1-38, 1989.
20. D.E. Rumelhart, F.E. Hinton and R.J. Williams, "Learning representations by back-propagation of errors", *Nature*, 323:533-536, 1986.
21. J.L. McLelland and D.E. Rumelhart, *Explorations In Parallel Distributed Processing*, MIT Press, Cambridge, Mass., 1989.
22. S. Haykin, *Neural Networks: a Comprehensive Foundation*, Macmillan College Publishing Co., N.Y., 1994.
23. SAS Institute Inc., *SAS User's Guide: Basics*, Version 5 edition, Cary, NC, 1985.
24. SAS Institute Inc., *SAS User's Guide: Statistics*, Version 5 edition, Cary, NC, 1985.
25. G.K. Bhattacharyya and R.A. Johnson, *Statistical Concepts and Methods*, Wiley, N.Y., 1977.
26. T. Ashiya and M. Nakagawa, "A proposal for a recognition system for the species of birds receiving birdcalls - an application of recognition systems for environmental sound", *IEICE Trans. Fundamentals*, E76-A:1858-1860, 1993.
27. R.A. Jacobs, M.I. Jordan, S.J. Nowlan and G.E. Hinton, "Adaptive Mixtures of Local Experts", *Neural Computation*, 3:79-87, 1991.