FEATURE SELECTION USING MUTUAL INFORMATION AND STATISTICAL TECHNIQUES

Xiangyu Sally Song Janusz A. Starzyk
School of Electrical Engineering & Computer Science, Ohio University

ABSTRACT NUMBER: 499

SESSION: Adaptive Signal Processing Techniques

CONTACT INFORMATION:
Xiangyu Sally Song
4540 Risinghill DR.
Plano, TX, 75024
Email: xiangyu_song@yahoo.com

AUTHOR AND CO-AUTHER AFFILIATION
Dr. Xiangyu Sally Song, Ph.D., School of Electrical Engineering and Computer Science of Ohio University.
Dr. Janusz A. Starzyk, Professor of Electrical Engineering and Computer Science of Ohio University, Athens, OH.

SENT TO:
MILCOM 2001 Security
The MITRE Corporation
7515 Colshire Dr.
Mail Stop N120
McLean, VA 22102

FEATURE SELECTION USING MUTUAL INFORMATION AND STATISTICAL TECHNIQUES (U)

Xiangyu Sally Song Janusz A. Starzyk
School of Electrical Engineering & Computer Science,
Ohio University, Athens, OH

Date of Submission June 1, 2001

(U) ABSTRACT

(U) This paper develops a feature selection method for automatic target recognition using Mutual Information (MI) approach and Statistical Techniques. Using Mutual Information criteria, an efficient feature selection method (repetitive transformation on the initial set of features, selective combination of partitioned feature subspaces) was developed based on HRR data. The developed procedures yield a minimum set of the most informative features. A statistical analysis on the input space and partitioned feature space was made based on a developed confidence interval model for mutual information. By incorporating the confidence interval of mutual information, the feature selection procedures developed in [1] were modified in order to make the highest increase of the lower bound mutual information in each feature's generation. As a result, the overall recognition rate will increase based on more reliable features, and the feature generation procedures involve less expensive hardware implementation. Moreover, a new input data acquisition procedure in order to optimize the construction of the training data base can be developed using the developed feature selection and evaluating procedures in this paper. Eventually, the whole automatic target recognition system can be implemented on an ontogenic neural network, which has self-organizing ability to accommodate new knowledge with only local alternations to its structure.

I. (U) INTRODUCTION

(U) During the development of neural net classifiers the "preprocessing" stage, where an appropriate number of relevant features is extracted from the raw data, has a crucial impact both on the complexity of the learning phase and on the achievable recognition performance. While it is essential that the information contained in the input vector is sufficient to determine the output class, the presence of too many input features can burden the training process and can produce a neural network with more connection weights than those required by the problem. From an application point of view, an excessive input dimensionality implies lengthened preprocessing and recognition times, even if the learning and recognition performance is satisfactory. The proposal of using Mutual Information as a criterion for selecting features is to limit the input dimensionality, and the analysis based on mutual information provides a useful diagnosis of the relevance of different features and of mutual dependencies. Roberto Battiti [2] investigated this topic and proposed an algorithm that is based on a "greedy" selection of the features and that takes both the mutual information with respect to the output class and with respect to the already-selected features into account. Other relevant work can be referred to [3,4,5].

II. (U) A HARDWARE EFFICIENT FEATURE SELECTION BASED ON MUTUAL INFORMATION-REPETITIVE TRANSFORMATION (RT) METHOD

(U) In searching for a feature selecting method that will give a minimum set of the most informative features and is hardware implementable, we developed a repetitive transformation procedure. In the view of feature space, each feature divides the sample space into two feature subspaces. Sequentially, N features will generate 2^N feature subspaces without combining any of the subspaces. But with some constraints, feature subspaces can be combined on purpose (see "Combination Method" in III), accordingly, the logic combination of feature is also need to be rearranged when input to classifier, so that the total feature subspaces can be less than 2^N. MI is defined in (2.1) with respect to the training classes (C) and the feature space (F) partitioned by the feature already been selected.

$$\begin{aligned} I(C;F) = I(F;C) &= \sum_{c,f} p(c,f)\log\frac{p(c,f)}{p(c)p(f)} \\ &= \sum_{c,f} p_{c,f}\log(p_{c,f}) - \sum_{f} p_f \log(p_f) - \sum_{c} p_c \log(p_c) \end{aligned} \quad (2.1)$$

(U) First, a limited subset of relevant features are selected from the initial set of available features. This is a Sequential Selection where the next feature selected is the feature from the initial set of available features that will give the most increase in MI with respect to the training classes and the feature space partitioned by the feature already been selected. This selection ends if the increase of MI is lower than a threshold. Then a Wavelet Transformation (WT) is iteratively applied to the subset of the selected features and Sequential Selection is applied to these transformed features. The process is continued in this way, using a limited number of iterations, until a set of the most informative feature is selected. This procedure is done before selected features are fed to the neural net, so the processing time is independent on the training process. By using the same wavelet transformation repetitively, hardware implementation can be greatly simplified.

(U) Experiments demonstrated that a limited iteration can generate the most informative features, i.e., the set of features that give the largest MI. The experiment was done on a set of training data shown in Fig.1. It contains equal number of three class objects, represented by different gray levels. The x-axis and y-axis stand for the signal magnitude received for a airplane target from two different directions.

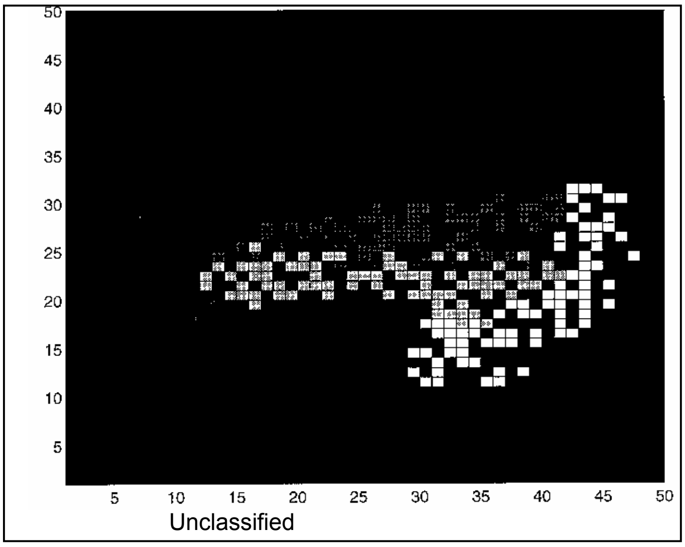

Fig.1. (U) Training Data

(U) The initial feature set consists of 18 features including signal values, maximum signal value, mean signal value, standard deviation, Shannon entropy, lp norm etc. The repetitive WT is Harr WT on the first four features selected in each iteration. Fig.2 is the MI verses features sequentially selected from the initial feature set. Fig.3 is the MI verses the features generated using RT. We can see from Fig.3, after four iterations, the MI reaches high values.

III. (U) CONFIDENCE INTERVAL IN MUTUAL INFORMATION

(U) Because MI is calculated by estimating the probability density from a finite number of samples, we must take into account the errors caused by the estimation. And from there, we develop the concept of Confidence Interval for MI, and then develop a feature selecting procedure that sequentially selects features based on the increase of the lower bond of the MI. In order to reduce the MI error, we develop a method called "Combination Method" which under some constraints combines some feature subspaces. This method produces a set of more reliable features under a certain confidence level than the feature sets generated based only on the increase of MI.

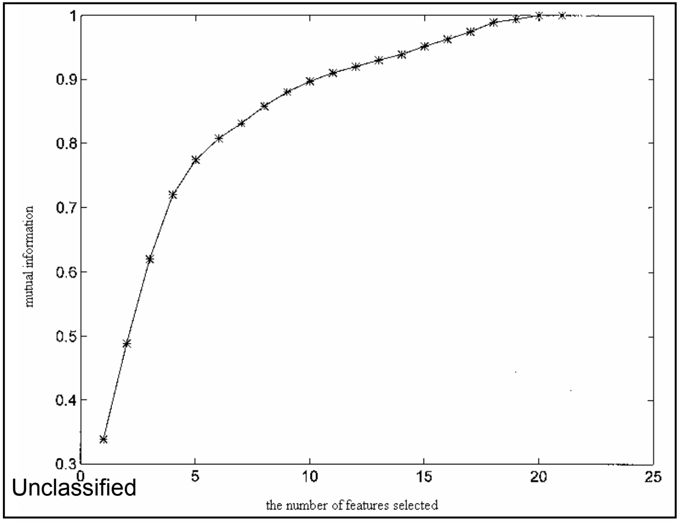

Fig.2. (U) MI verse features generated sequentially

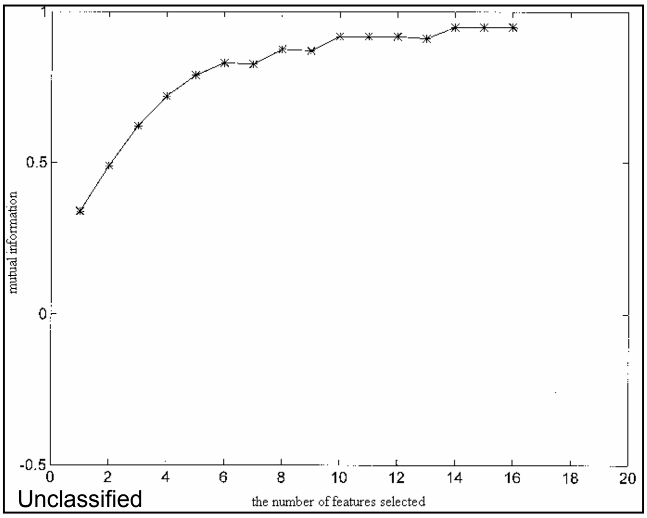

Fig.3. (U) MI verses features generated using RT method

A. (U) STATISTICAL ANALYSIS OF PROBABILITY AND ITS CONFIDENCE INTERVALS

(U) In MI definition, probability is calculated as proportion $p = x / n$, n is the population, x is the sample count. In statistical theory, n trials satisfy the assumptions of the Binomial distribution and the sample proportion is an unbiased estimation of the true proportion p to be estimated on the basis of a sample. When $np \le 10 (n \ge 100)$, Binomial distribution is best approximated with a Poisson distribution with $\lambda = np$. The upper estimation for p is $p < \frac{1}{2}\mathrm{N}_\alpha^2$. N_α^2 is the area under the Chi-square distribution to whose right equal to α. When $np > 5$ and $n(1-p) > 5$, Normal distribution provides good approximation to the Binomial distribution. Given confidence (1-α) 100%, the confidence limits for p are:

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}} \qquad (3.1)$$

(U) Based on these observations, we use polynomial approximation to approximate p in the range that is not defined above up to first differential continuity. And use $a\tan()$ function to limit the upper bound of p to 1 when doing approximation. Thus we get the lower and upper bound for p, which is named as confidence interval for p. See Fig. 4. The innermost line is for 100 points population and the outermost line is for 1000 population.

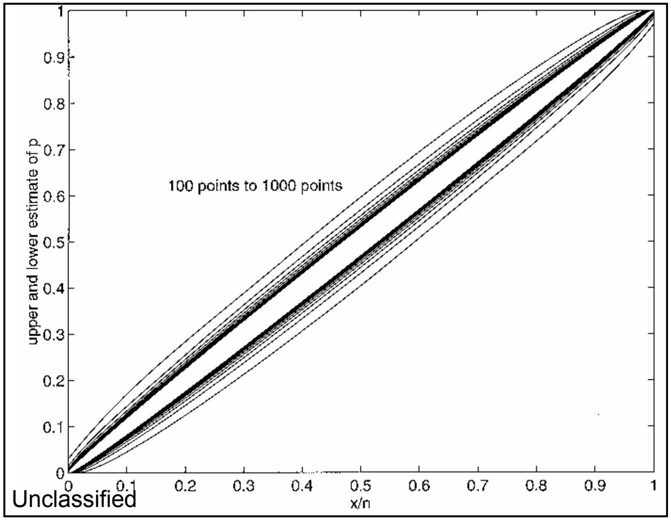

Fig.4. (U) Upper and lower estimate of p for different populations

B. (U) MUTUAL INFORMATION ERROR

(U) Given the confidence interval for p, we can calculate the error involved in calculating $p(\log(p))$ and denote it as $err(p(\log(p)))$. We define it to be the maximum difference among values of $p(\log(p))$ when p changes its value within its confidence interval. The mutual information error is calculated as $Inferr = \sum_i err^2(p(\log(p))$, where the summation is over all the $p(\log(p))$ items in the mutual information definition. Then the lower bound for the mutual information is Lowinf=I-Inferr. To illustrate this, some experimental results are shown here. In order to give a visible illustration, we dealt with a two-class problem and assume there is a feature dividing the feature space into two, Fig.5. to Fig.7 show how the distribution of these two class sample count in these two subspaces affects the distribution of information error and Lowinf.

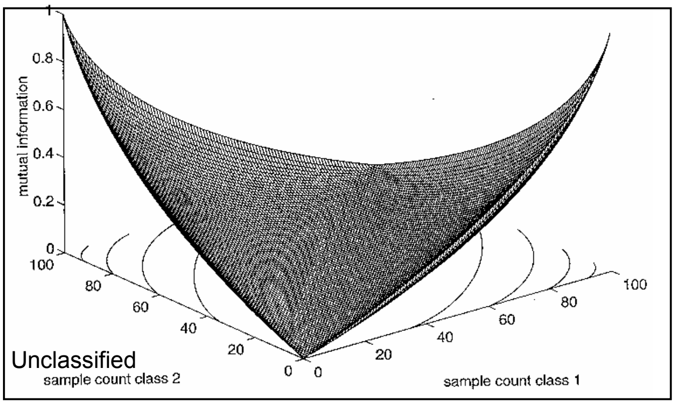

Fig.5. (U) MI verses sample distribution

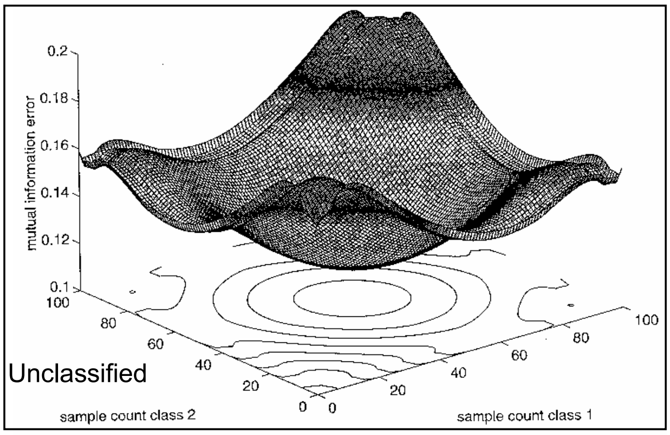

Fig.6. (U) Inferr verses sample distribution

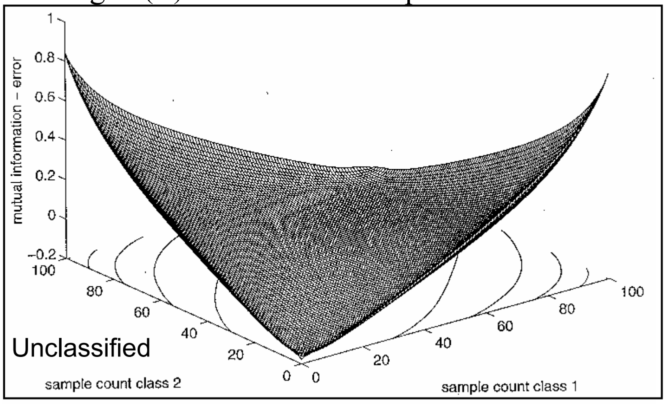

Fig.7. (U) Lowinf verses sample distribution

(U) Due to a certain confidence interval which exists for every sample estimate of p, the mutual information also has uncertainty involved, expressed in Inferr. And the Lowinf is the mutual information we can at least get under the pre-assigned confidence degree.

C. (U) COMBINATION METHOD FOR SELECTING FEATURES

(U) At this point, we believe to build a more reliable system, the Lowinf should be used instead of Mutual Information, the feature selection procedure should achieve the Lowinf increase when each feature is selected. In the method previously developed, the MI increases with each generated feature, but the Inferr also increases due to the increase of the number of the subspaces. Actually, the Lowinf decreases after several features were selected. See Fig.8.

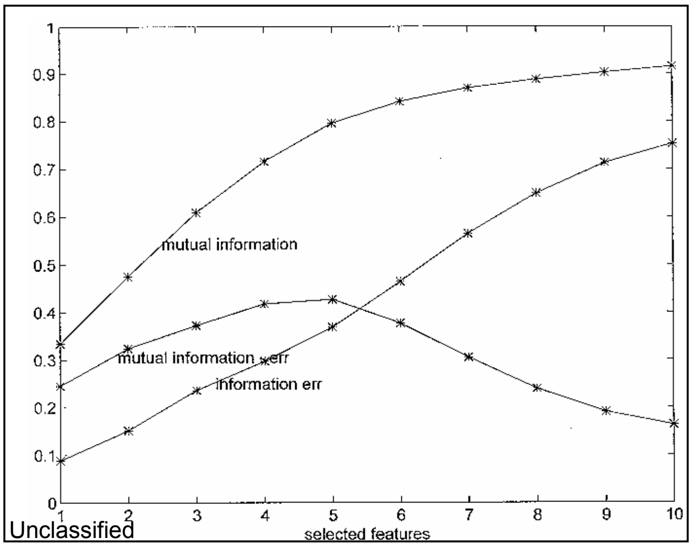

Fig.8. (U) MI, Inferr and lowinf verses features selected

(U) In order to reduce the Inferr and obtain as large Lowinf increase as possible with each generated feature, a method called "combination method" was developed. The procedure is as follows:

1. (U) Select an initial feature set .The initial feature space is the subspace contains all the training data. A subset of features will be sequentially selected in step 2 and step 3 from the initial feature set.

2. (U) For the q_{th} feature selection: First for all the features in the initial feature set, find the best partition point of each feature. A partition point means a certain value for this feature. There should be the same amount of partition points as the amount of training data. The best partition is the one that can give the largest Lowinf increase. Here for each partition, we use "selective combination of subspace" to reduce the amount of subspaces, thus, reduce the Inferr and gain Lowinf as large as possible. Now we have q features and m ($m \leq q^2$) feature subspaces denoted as 1, 2 ...i ... m. The Lowinf for subspace i is denoted as Lowinf(i) Then for a certain partition of the next feature, each subspace i should be divided into 2 subspaces, i_1 and i_2, without combining any subspace. If Lowinf (subspace i1 and i2) > Lowinf(subspace i), we accept this division. If not, we combine subspaces i_1 and i_2 back to one subspace. And the Lowinf for this partition is calculated based on the subspaces that have been partly combined under the condition listed above. Record best partition for each feature and the Lowinf at this partition, select the feature which has the largest Lowinf as the candidate feature and record its best partition and how the subspaces look like after this feature be selected under this particular partition value. Calculate the increase in Lowinf as the difference between the Lowinfinc now and that before this feature being selected).

3. (U) If Lowinfinc> threshold, this candidate feature is selected. The whole feature space is updated under this feature's best partition and the combination of some subspace pair into one. The updated feature space will be used for the $(q+1)_{th}$ feature selection. Repeat 2 and 3, until the Lowinfinc< threshold. At the end, we get a set of selected features that give the largest Lowinf .

D. (U) SIMULATION AND COMPARISON

(U) Simulations are done to demonstrate the "Combination Method". Comparison is made between this method and a reference method--"Non-combination Method". The steps for reference feature selection method are the same as in III.C, except that the feature is generated based on the MI instead of Lowinf as in "Combination Method" and no combination of subspaces involved. The training data and the initial feature set are the same as those in II. The results are shown in Fig.9. to Fig. 11.

(U) It can be seen that the "Combination Method" stops when there is no increase in Lowinf while "Non-combination Method" stops when there is no increase in MI. Even through "Non-combination Method" can select more features to reach higher MI than the "Combination Method", but after several features, the error in MI for the "Non-combination Method" is very large, actually, it can be seen that its Lowinf decreases after a certain number of features been selected, that means the following features are not reliable according to our confidence degree. It can be seen that the "Combination Method " can reach much higher Lowinf than the "Non-combination Method".

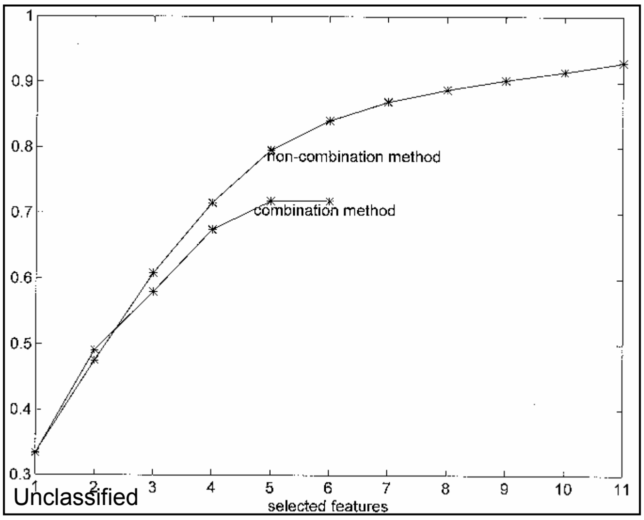

Fig.9. (U) Mutual information comparison

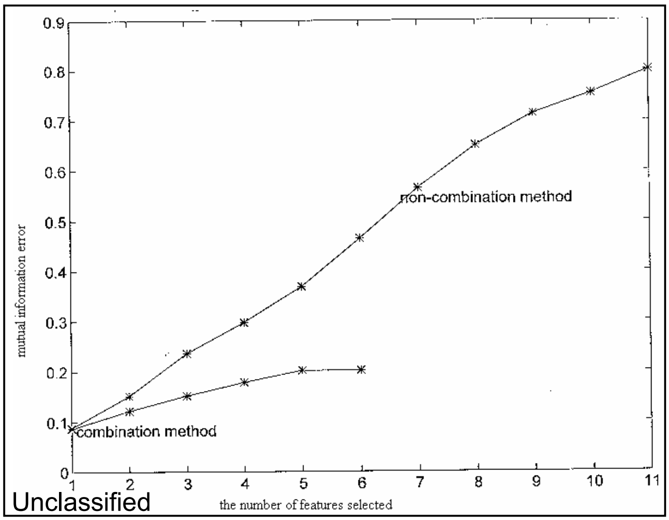

Fig.10. (U) Mutual information error comparison

IV. (U) CONCLUSIONS

(U) This paper employs Mutual Information and Statistical Analysis Method to select features. It developed a Repetitive Transformation Method to generate features in order to gain hardware simplicity. Also, the concept of confidence interval in statistical analysis theory has been used in Mutual Information to lead to the use of Lower bound of Mutual Information as a criterion other than Mutual Information to generate more reliable features. As a result, the overall recognition rate will increase based on more reliable features, and the feature generation procedures involve less expensive hardware implementation. The developed feature selection and evaluating procedures can be used to develop a new data acquisition process to optimize the construction of the training database. Training database is usually very expensive to acquire. This procesure will start from a minimum set of training data and request new ones only when needed based on a certain level of mutual information and confidence degree. Eventually, the whole automatic target recognition system can be implemented on an ontogenic neural network which has self-organizing ability to accommodate new knowledge with only local alternations to its structure.

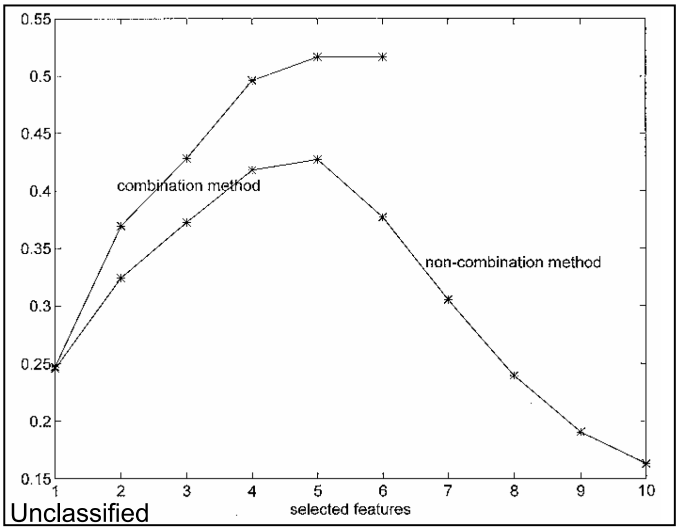

Fig.11. (U) Mutual information-err (Lowinf) comparison

REFERENCES

[1] Janusz A. Starzyk, " Feature selection for ATR neural network approach," Final report for Summer Faculty Research Program, August, 1996, Wright Laboratory.

[2] Roberto Battiti, " Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Networks, vol.5, No.4, July 1991, pp.537-550.

[3] M.Bichsel and P.Seitz, "Mimimum class entropy: A maximum information approach to layered networks," Neural Networks, 2:133-141, 1989.

[4] J.S.Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in, Advances in neural Information Processing Systems, vol.2, D.S.Touretzky, ed. San Mateo, CA: Morgan Kaufmann, 1990, pp.211-217.

[5] F.Kanaya and K.Nakagawa, "On the practical implication of mutual information for statistical decision making," IEEE Trans. Information Theory, vol.37, pp.1151-1156, 1991.