International Journal of Signal System Control and Engineering Application 2 (1): 30-34, 2009
ISSN: 1997-5422

Pattern Classification Using a Fusion of the Infomax and Imax Algorithms

Mohamed Deriche
Department of Electrical Engineering, King Fahd University of Petroleum and Minerals,
Dhahran 31261, Saudi Arabia

Abstract: A new algorithm for feature selection based on information maximization is derived. This algorithm performs subspace mapping from multi-channel signals, where Network Modules (NM) are used to perform the mapping for each of the channels. The algorithm is based on maximizing the Mutual Information (MI) between input and output units of each NM and between output units of different NMs. Such formulation leads to substantial redundancy reduction in output units, in addition to extraction of higher order features from input units that exhibit coherence across time and/or space useful in classification problems. We discuss the performance of the proposed algorithm using two scenarios, one dealing with the classification of EEG data while, the second is a speech application dealing with digit classification.

Key words: Mutual information, infomax, imax, PCA, CCA, feature selection

INTRODUCTION

The problem of subspace mapping has attracted a lot of attention among researchers. One of the well established solutions to the problem is Principal Component Analysis (PCA), which maps correlated input units into uncorrelated output units (Devijver and Kittler, 1982). Canonical Correlation Analysis (CCA) (Mardia *et al*., 1979), on the other hand, has been proposed as an alternative for the case of two variable sets and aims at maximizing correlation between output units by linearly transforming inputs.

Many Artificial Neural Networks (ANNs) have been developed to implement PCA (Karhunen and Joutesnsalo, 1995; Miao and Hua, 1998; Al-Ani and Deriche, 2001; Torkkola, 2003) and CCA (Lai and Fyfe, 1999). However, no attempt has been made to combine the power of the two techniques. There is a clear advantage in combining PCA and CCA, then generalizing the approach to multiple channels. To achieve this, we propose an information theoretic based approach, which we name the Information Maximization Feature Selection (IMFS) algorithm. The approach is based on combining the infomax algorithm proposed by Linsker (1988, 1997, 2005) and the Imax algorithm developed by Becker (1996), Slonim and Weiss (2003), Torkkola (2003) and Agakov and Barber (2005). Two applications (EEG, data, speech data) to evaluate the performance of the proposed algorithm.

Correlation based techniques: One of the most widely used dimension reduction techniques is Principal Component Analysis (PCA) (Devijver and Kittler, 1982). Starting with an input x of N units, PCA finds the directions along, which the variance is maximal. If .>...>. are the first k eigenvalues of the covariance matrix, R, then the corresponding eigenvectors a,...,a, become the first k principal components of R. It can be proven that PCA is an optimal linear dimension reduction technique in mean-square sense, which ensures that output units are uncorrelated.

For the cases of two variable sets, CCA has been developed to summarize relationship between the output units by finding a linear combination, for each set, which results in the highest correlation between the output sets. If inputs x and y are of dimension N, then CCA (Mardia *et al*., 1979) finds the two M dimensional vectors, such that the correlation between these is maximal.

Information based techniques: The infomax principle was first developed by Linsker (1988) and was inspired from Hebb's rule: if unit a is one of the input units contributing to output unit b and if a tends to agree with b, then the future contribution that the firing of a makes to that of b is increased. In other words, connection strengths are modifying according to the degree of dependency between input and output, which acts to generate an output unit whose output activity preserves maximum information about the input activity, subject to constraints.

For the case of two channels, Becker (1996) proposed the Imax algorithm, which was inspired from human sensory processing. A major feature of the sensory data is coherence across time and across different sensory channels, where coherence means that one part of the

signal can somehow be predicted from another part. It has been argued that spatio-temporal and multi-sensory coherence provides important cues for representing signals in space and time and are useful in object localization and identification. The main idea is that two different NMs can learn to extract features that are coherent across their inputs.

MATERIALS AND METHODS

The IMFS algorithm

Maximizing MI within a NM: Maximizing MI between the input and output units of a single NM is equivalent to maximizing the total information conveyed by the output units and minimizing the information that the output units convey to someone, who has prior knowledge about the input units. In this study, we consider the case, where we want to map an (N×1) input, x, into an (M×1) output, p, with M<N, p = Wx+n. Where, n is additive noise, uncorrelated with x. Using basic concepts from information theory and assuming that both input and noise are Gaussian, the MI becomes:

$$\begin{aligned} I(p, x) &= h(p) + h(x) - h(p, x) \\ &= h(p) - h(n) \\ &= 0.5 \log [|R_{pp}|] - 0.5 \log[|R_{nn}|] \end{aligned} \quad (1)$$

where:

h(p) = The entropy

R = The covariance matrix of p

$$R = WR_{xx}W + R_{nn}$$

Maximizing MI between x and p can be achieved using the learning rule:

$$\begin{aligned} \frac{\partial I(p;x)}{\partial W} &= 0.5\frac{\partial \log[|R_{pp}|]}{\partial W} \\ &= (WR_{xx}W^T + R_{nn})^{-1}WR_{xx} \end{aligned} \quad (2)$$

The optimal value for M can be set based on the amount of information lost in the output units in a similar way to retention of eigen values in PCA.

Maximizing MI between the output units of two different NMs: Unlike the Imax algorithm that considers the two observed signals at the output of each NM as noisy versions of the same underlying signal, p = x + n, q = x + m, we consider here that two different signals are used as input units to each NM and the objective is to transform these linearly to the output units such that the MI between the output units of the two NMs is maximized, p = Wx + n, q = Vy + m, where W and V are the weights of the two NMs. Let a = [p q], then (Fig. 1):

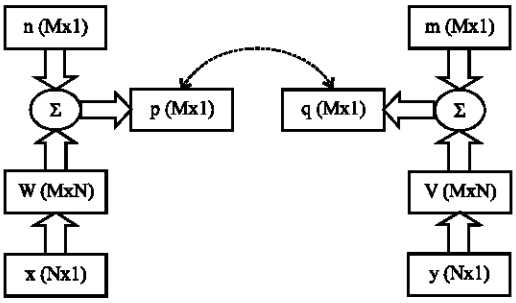

Fig. 1: Proposed MIFS algorithm

$$\begin{aligned} I(p,q) &= h(p) + h(q) - h(p,q) \\ &= 0.5\log[(2\pi e)^M |R_{pp}|] \\ &+ 0.5\log[(2\pi e)^M |R_{qq}|] - 0.5\log[(2\pi e)^{2M} |R_a|] \\ &= 0.5\log[|R_{qq}|/|R_{qq}R_{qp}R_{pp}^{-1}R_{pq}|] \end{aligned} \quad (3)$$

$$\begin{aligned} R_a &= \begin{bmatrix} R_{pp} & R_{pq} \\ R_{qp} & R_{qq} \end{bmatrix} \\ &= \begin{bmatrix} WR_{xx}W^T + R_{nn} & WR_{xy}V^T \\ VR_{yx}W^T & VR_{yy}V^T + R_{mm} \end{bmatrix} \end{aligned} \quad (4)$$

Maximizing MI between p and q can be achieved by updating W, (respectively V) according to the learning rule:

$$\begin{aligned} \frac{\partial I(p;q)}{\partial W} &= R_{pp}^{-1}R_{pq}(R_{qq} - R_{qp}R_{pp}^{-1}R_{pq})^{-1} \\ &\times [VR_{yx} - R_{qp}R_{pp}^{-1}WR_{xx}] \end{aligned}$$

and

$$W = W + \alpha\frac{\partial I(p;q)}{\partial W} \quad (5)$$

The maximization is constrained to a normalized W (|WW = 1|), which can be achieved using Lagrange multipliers.

Extension to multiple channels: Equation 2 and 5 were first re-derived for the case of three channels with closed form expressions obtained for updating the weighting matrices for each of the three Channels. However, the derivation led to complex expressions for I(p; q; r)/U, (respectively for W and V). To reduce the complexity resulting from the above, we propose here a fast approximation using pairs of channels. In the case of 4 channels (p, q, r, s), for example, the updating rule for channel 1 (W) becomes:

$$W = W + \alpha \left[\frac{\partial I(p;q)}{\partial W} + \frac{\partial I(p;r)}{\partial W} + \frac{\partial I(p;s)}{\partial W} + \frac{\partial I(q;r)}{\partial W} + \frac{\partial I(q;s)}{\partial W} + \frac{\partial I(r;s)}{\partial W} \right] + \beta \frac{\partial I(p;x)}{\partial W} \quad (6)$$

where, . and . are the learning rates used to weight the outer and inner information loss. Even though, the Eq. 6 is not optimal, the extensive simulations have shown that most of the MI within channels is preserved, while, still maintaining a reasonable amount of information between channels.

RESULTS AND DISCUSSION

Choosing appropriate learning rates: We carried experiments on both speech (classification) and EEG data. Here, we present the results from analyzing EEG data from neighboring channels of an 8 sec segment EEG data that represent the left and right movements. The 13 features were estimated from the data and used as input to the two NMs. The features were dominant frequency and its amplitude, average power in main lobe, energy, zero crossing and number of extreme of each segment, average half-waves amplitude and duration and poles of AR model.

Figure 2a and b display the effect of . and . on updating W to maximize MI within the two NMs and between their output units versus the true values of I (p; x) and I (p; q).

From Fig. 2a and b, we notice that by only updating the weights to maximize MI between the two NMs, we cannot guarantee maximum MI within each one of NMs and vice versa. Even though, an optimal choice for . and . is application dependent, we found that an initial choice of . = 0.8 and n(n-1)/2. = 0.2 (where, n(n-1)/2 is the number of channel pairs) leads to a minimum information preservation of 90% or more (fine tuning these parameters is possible depending on the application of interest).

It is worth mentioning that the computational cost involved is not significant, where adjusting the weight matrices for 100 iterations using Matlab routine running under conventional PC environment takes <1 second (3 matrix inversions of M×M at most).

Comparison to PCA and CCA: We first carried some initial experiments with two channels. By varying the values of . and . between 0 and 1, we could see the IMFS converging towards either the PCA (. = 0) or the CCA (. = 0). However, in contrast with PCA and CCA, the IMFS is able to maximize both I(p; x) and I(p; x) at the same time (with <10% in information loss).

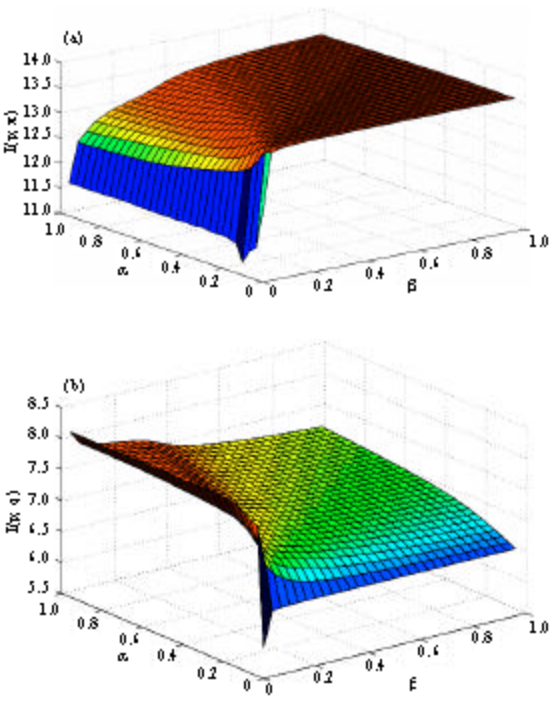

Fig. 2: MI between a): Input set x and output set p and b): Output sets p and q, using 2-channel EEG data

We then carried an experiment using five channels of EEG and compared the performance of the approximate IMFS using subsets of 2 and 3 channels with that of PCA (Fig. 3a, b).

Even though, these two versions of the IMFS only achieve near optimal results for I(p; q; r; s; t), they still outperform, by far, PCA. In addition, the IMFS can be used for any number of channels without the need to derive complex formulas for the learning rules, while achieving very promising results.

A second experiment in speech classification based on 2-channel speech data generated by artificially adding white Gaussian noise to the original data obtained from the TIMIT database. The experiment consists of classifying the 10 digits. The features obtained by applying PCA, CCA and HIM with the number of output elements ranging from 1-40 were used to represent speech segments that are fed to an artificial neural network. The features used include Linear Prediction Coefficients (LPC), Line Spectrum Pair (LSP), Reflection Coefficient (RC), Cepstrum Coefficients (CC), Mel Frequency Cepstrum Coefficients (MFCC), Filter Bank Coefficients (FB) and Wavelet Coefficients (WC).

Figure 3b shows the average classification accuracy of two simulated speech channels for different number of output elements. It can be shown from Fig. 3b that the performance of the PCA is better than that of CCA

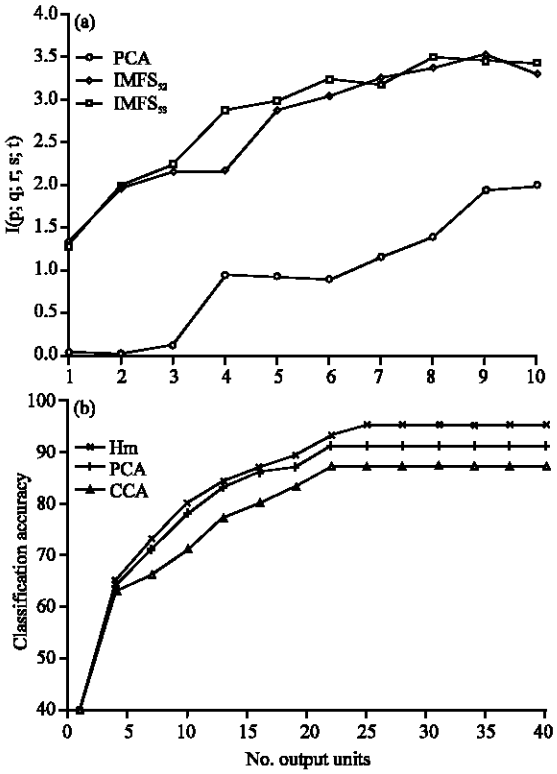

Fig. 3: Comparison between PCA, CCA and HIM, a): . = 1.0, . = 0.0 using 2-channel EEG data and b): . = 0.0, . = 1.0 using 2-channel speech data

(Campernolle *et al*., 2005). The proposed algorithm, HIM, on the other hand, approaches the PCA when M is small, then it outperforms PCA when there are >20 elements. The performance obtained was better the results discussed in similar earlier research (Campernolle *et al*., 2005).

Several other experiments were carried to test the power of the algorithm in the presence of noise and for the case of the generalized Gaussian distributions with very promising results.

CONCLUSION

A new feature extraction algorithm based on the maximization of MI between the output units of different NMs as well as between the input and output units of individual NMs has been developed. We have shown that by assigning appropriate learning rates to the two cost functions, a new cost function that preserves information content, between the two NMs and within each NM, can be obtained.

The initial experimental results using synthetic and real EEG and speech data showed clearly the power of the proposed algorithm as compared to PCA and CCA and the previously developed Infomax and Imax algorithms. The concept proposed here is novel with a great potential in optimal feature extraction from multi-channel data using information theory concepts.

ACKNOWLEDGEMENTS

The author would like to thank King Fahd University of Petroleum and Minerals (KFUPM) and King Abdulaziz City for Science and Technology (KACST), Saudi Arabia, for supporting the research work discussed in this study.

REFERENCES

Agakov, F. and D. Barber, 2005. Variational information maximization for neural coding. Lecture Notes in Computer Science, LNCS, 3316/2004: 543-548, DOI: 10.1007/b103766, http://www.springerlink.com/content/ dg2833x65kger3ag.

Al-Ani, A. and M. Deriche, 2001. A dempster-shafer theory of evidence approach for combining trained neural networks. The 2001 IEEE Int. Symposium on Circuits and Syst. ISCAS, 2, 3: 703-706. DOI: 10.1109/ISCAS.2001.921429, ieeexplore.ieee.org/iel5/7344/19927/00921429.pdf In ISCAS'2001.

Becker, S., 1996. Mutual information maximization: Models of cortical self organization. Network: Computation in Neural Syst., 7 (2): 7-31. (electronic) 0954-898X (paper). http://www.informaworld.com/smpp/title~db=all~content=t713663148.

Campernolle, D., R. Cools, M. Matton and M. Wachter, 2005. Maximum mutual information training of distance measures for template based speech recognition. Proc. International Conference on Speech and Computer, pp: 511-514. www.esat.kuleuven.be/psi/spraak/cgi-bin/get_file.cgi?/mmatton/specom05/paper/paper.pdf.

Devijver, P.A. and J. Kittler, 1982. Pattern recognition: A statistical approach. Prentice-Hall, ISBN: 10-0136542360. http://personal.ee.surrey.ac.uk/Personal/J.Kittler/cv.html.

Karhunen, J. and J. Joutesnsalo, 1995. Generalizations of principal component analysis, optimization problems and neural networks. Neural Networks, 8 (4): 549-562. DOI: 10.1016/0893-6080(94)00098-7. http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T08-4031CR2-1S&_user=1074406&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000051301&_version=1&_urlVersion=0&_userid=1074406&md5=7132cf557384c7a245a7f186c6daa1bb.

Lai, P.L. and C. Fyfe, 1999. A neural implementation of canonical correlation analysis. Neural Networks, 12(10): 1391-1397. DOI: 10.1016/S0893-6080(99)00075. http://www.sciencedirect.com/science/article/B6T08-3XXCXC7-5/2/5d45c4fe10450c5fb87042a19c4e6502.

Linsker, R., 1988. Self-organization in perceptual network. Computer, 21 (3): 105-117. DOI: 10.1109/2.36. http://portal.acm.org/citation.cfm?id=47869#.

Linsker, R., 1997. A local learning rule that enables information maximization for arbitrary input distributions. Neural Computation, 9 (8): 1661-1665. DOI: 10.1162/neco.1997.9.8.1661. http://portal.acm.org/citation.cfm?id=1246445.

Linsker, R., 2005. Improved local learning rule for information maximization and related applications. Neural Networks, 18 (3): 261-265. DOI: 10.1016/j.neunet.2005.01.002. http://www.sciencedirect.com/science/article/B6T08-4FV35JG-1/2/c4d87e5f5a16f78ddf591dceaa595c08.

Mardia, K.V., J. Ken and J. Bibby, 1979. Multivariate analysis. Academic Press. DOI: 10.1002/bimj.47102-40520. http://www3.interscience.wiley.com/journal/114077456/abstract?CRETRY=1&SRETRY=0.

Miao, Y. and Y. Hua, 1998. Fast subspace tracking and neural network learning by a novel information criterion. IEEE. Trans. Signal Process., 46: 1967-1979. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00700968.

Slonim, N. and Y. Weiss, 2003. Maximum likelihood and the information bottleneck. Adv. Neural Inform. Process. Syst., 15: 335-342. http://www.cs.huji.ac.il/~yweiss/MLandIB7.ps.

Torkkola, K., 2003. Learning boolean concepts in the presence of many irrelevant features. J. Machine Learning Res., 3: 1415-1438. http://jmlr.csail.mit.edu/papers/volume3/torkkola03a/torkkola03a.pdf.