International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Pattern Recognition using Mixture of Experts - A Review

Pramod P Nair

Department of Mathematics, School of Arts and Sciences,
Amrita Vishwa Vidyapeetham, Amritapuri, (India)

ABSTRACT

The present and possible future roles of Mixture of Experts and Hierarchical Mixture of Experts models, especially for pattern recognition is discussed and reviewed. Both supervised and unsupervised techniques are explored along with the various techniques for parameter estimation. The EM algorithm, which forms a vital part in most of the parameter estimation algorithms, is discussed in detail and its various forms of maximization procedures. A synthesis is made of the unresolved problems related to its application in image classification. Finally an outlook into the future applications of ME and HME has been presented.

Keywords: Expectation Maximization (EM) algorithm, Generalized linear model (GLIM), Hierarchical Mixture of Experts (HME), Multiclass Classification, Mixture of Experts (ME).

I.INTRODUCTION

One of the important modules of early vision problem is image segmentation and classification. Segmentation can be defined as the process of partitioning an image into some distinct homogeneous regions with respect to the characteristics measured such as intensity, color or texture. No two such regions are similar with respect to these characteristics. Classification is the task of labeling the segments where segments with same characteristics are given similar labels. Numerous segmentation and classification techniques are available in literature and they have been mentioned and explored in various surveys [1] [2] [3] [4].

In the past twenty years, the study of neural networks has been enriched by an infusion of ideas from diverse fields, including statistics and probability theory, information theory, physics, and biology. In the mid-eighties, the back-propagation learning algorithm for neural networks was introduced. This algorithm for the first time made it feasible to train a non-linear neural network equipped with layers of the so-called hidden nodes. Since then, neural networks with one or more hidden layers can, theoretically, be trained to perform virtually any regression or discrimination task. Moreover, no assumptions are made as with respect to the type of underlying (parametric) distribution of the input variables, which may be nominal, ordinal, real or any combination hereof.

From the early nineties, ANN [5] [6] [7], Radial basis functions [8], Hopfield networks [9] were widely used for both pixel-based and feature-based image segmentation. The idea to incorporate the consensus rule and obtain better classification led to structure of committee machines [10] ,a neural network model that was inspired by mixture models from statistics [11], [12]. In 1991 Jacob and Jordan proposed the “mixture of experts” (ME) [13] method for classification. Three years later the “hierarchical mixture of experts” (HME) [14] was introduced by the same group of researchers. Since then ME’s have been suggested for a variety of problems, including classification [13], [15], [16] control [14], [17], and regression tasks [18], [19].

In this review, we survey the applications of ME and HME developed to solve different image classification problems. Here we shall try to find some applications of ME and HME in classification and discuss the major strengths and weaknesses of ME and HME for solving classification problems.

The basic theoretical concepts of ME and HME along with their probabilistic interpretations are proposed in section 2. Section 3 discusses some categorization and real world problems of these architectures. In section 4, the concluding section, the problems related to pattern recognition are considered and an overview of future research issues, which needed to be resolved or investigated further is presented.

II.THEORETICAL CONCEPTS

The basic principle used here is that of *divide and conquer*. In statistical and machine learning literature this principle has been widely used in well known algorithms such as the Classification and Regression Trees (CART) [20], Multivariate adaptive Regression Splines (MARS) [21], and Induction of Decision Trees (IDT) [22] much earlier. They fit by explicitly dividing the input space into nested sequence of regions. The notable factor is that their convergence in order of magnitude is faster than gradient based neural network algorithms.

Although the divide and conquer algorithms reduce the complexity and have favorable consequences for the bias of the estimator, it increases the variance. So as in most of divide and conquer algorithms, a piecewise constant or linear function is adopted to minimize variance at the cost of increased bias. A soft split of data is also allowed here, so that the input vector is allowed to belong to multiple regions simultaneously. The ME and HME are termed as dynamic committee machines. The term "dynamic" is used because the input is directly involved in actuating the mechanism that integrates the output of the individual experts into an overall output.

1.1. Mixture of Experts (ME) model

Let (**x**, *d*) be a training pattern where **x** is a vector of size M and *d* is the label to which the particular pattern belong. The network configuration of Fig.1 is referred to as mixture of expert model. It consists of K supervised modules called *Experts* and an integrating unit called the *gating network*. All the experts and the gating network are generalized linear models (GLIM) where the network is linear with single output non-linearity.

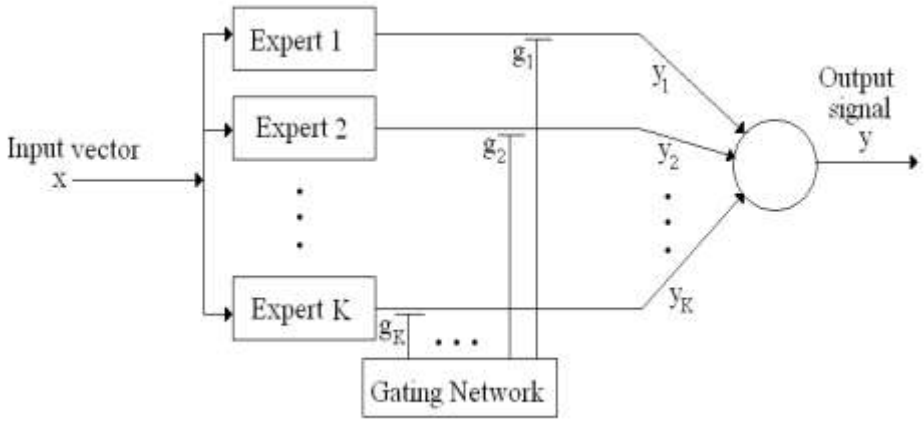

Fig1. Mixture of Experts architecture

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

The output produced by expert *k* is given by the log of the inner product of the input vector and the synaptic weight vector w_k of this neuron as:

$$y_k = w_k^T \mathbf{x}, k = 1,2,..K \quad (1)$$

The gating network consists of a single layer of K neurons, with each assigned to a specific expert. The activation function of the gating is defined as:

$$g_i = f(u_k), k = 1,2,..K \quad (2)$$

Where, u_k is the inner product of the input vector **x** and the synaptic weight vector a_k .

This activation function is referred to as softmax. Note that the g_i's are positive and sum to one for each **x**. They are interpreted as providing a soft partitioning of the input space. The overall output of the ME model is:

$$\mathbf{y} = \sum_{i=1}^{K} g_i y_i \quad (3)$$

Since both the *g*'s and the *y*'s depend on **x**, the total output is a non-linear function of the input.

1.2. Hierarchical Mixture of Experts (HME) model

The HME architecture (Fig. 2) is a tree in which the gating networks sit at the non-terminals of the tree. These expert networks receive the vector **x** as input and give outputs as in the case of ME model. The output vectors proceed up the tree, being blended by the gating network outputs.

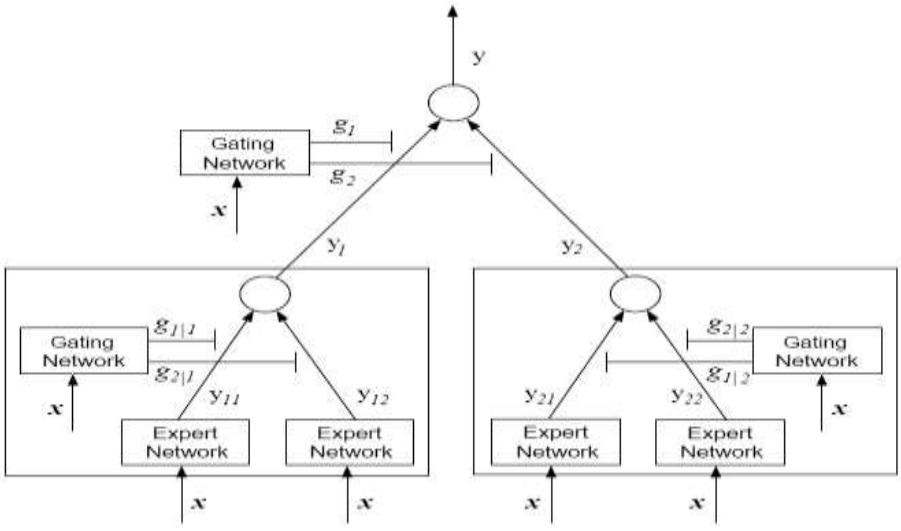

Fig.2 Hierarchical Mixture of Experts architecture

The gating networks are generalized linear systems. At the lower levels they are defined as follows:

$$g_{j|i} = f(u_{ij}) \quad (4)$$

Here $u_{ij} = a_{ij}\mathbf{x}$ is the output of the j^{th} unit in the i^{th} gating network at the second level of the architecture. The output at the i^{th} non-terminal is the weighted output of the experts below it. The output is given by:

$$y_i = \sum_j g_{j|i} y_{ij} \quad (5)$$

The top level gating is defined similar to gating of ME model and the overall output is calculated.

1.3. The Probability Model

Here we shall consider the probabilistic interpretation of the HME model. The interpretation of the ME model follows from this. A statistical model has been used to model the decision tree structures; in particular, the choice of parameterization corresponds to a *multinomial logit* probability model at each non-terminal node in [14]. A multinomial logit model is a special case of the GLIM that is commonly used for soft multiway classifications as in the case of self-organizing maps (SOM). Once a particular sequence of decisions has been made, resulting in a choice of regression process (*i, j*), the desired output *d* is assumed to be generated according to the following statistical model.

First, a linear predictor η_{ij} is formed:

$$\eta_{ij} = w_{ij}\mathbf{x} \tag{6}$$

The expected value of *d* is obtained by passing the linear predictor through a link function *f*:

$$y_{ij} = f(\eta_{ij}) \tag{7}$$

The output *d* is then chosen from a probability density *P*, with mean y_{ij} and dispersion parameter θ_{ij}. We denote the density of *d* as $P(d|\theta_{ij})$, where, the parameter vector θ_{ij} includes the weights y_{ij} and the dispersion parameter φ_{ij} .The density *P* is assumed to be a member of the exponential family of densities, particularly in [14], *P* is the *n*-dimensional Gaussian, and the dispersion parameter is taken as the identity matrix.

Given these assumptions, the total probability of generating d from **x** is the mixture of the probabilities of generating *d* from each of the component densities, where the mixing proportions are multinomial probabilities:

$$P(d|\mathbf{x},\theta) = \sum_i g_i \sum_j g_{j|i} P(d|\theta_{ij}) \tag{8}$$

Note that θ includes the expert network parameters θ_{ij} as well as the gating network parameters a_i and a_{ij}. Note also that we have explicitly indicated the dependence of the probabilities g_i and $g_{j|i}$ on the input **x** and on the parameters.

The terms *posterior* and *prior* have meaning in this context during the training of the system. The probabilities g_i and $g_{j|i}$ are referred as prior probabilities, because they are computed based only on the input **x**, without knowledge of the corresponding target output *d*. A posterior probability is defined once both the input and the target output are known. Using Bayes' rule, the posterior probabilities at the nodes are defined as follows:

$$h_i = \frac{g_i \sum_j g_{j|i} P(d|\theta_{ij})}{\sum_i g_i \sum_j g_{j|i} P(d|\theta_{ij})} \quad \text{and} \quad h_{j|i} = \frac{g_{j|i} P(d|\theta_{ij})}{\sum_j g_{j|i} P(d|\theta_{ij})} \tag{9}$$

III.CATEGORIZATION AND APPLICATIONS OF ME AND HME MODELS

The ME models are categorized on the part played by the gating network. Convectional ME [13] and their extensions [15], [16], [18], partition the problem space stochastically into subspaces and the experts are set to

specialize in each subspace using the gating networks. These networks are termed Mixture of implicitly localized experts (MILE). In another category termed the Mixture of explicitly localized experts (MELE), the problem space is explicitly partitioned by clustering methods and each expert is assigned one of the subspaces. The proposed MELE methods [23], [24] perform better than MILE since they partition the input space into more separable space. An extended comparison of these methods can be found in the literature survey by Saeed and Ebrahimpour [25].

The applications of ME and HME models include mainly pattern recognition tasks and regression problems. In pattern recognition, these models are widely used in signal processing [26] [27], speech recognition [28], and image classifications [29], [15], [16]. A supervised version of the EM algorithm using IRLS algorithm in the inner loop has been discussed in [14]. Other convectional methods like the gradient descent [10], Newton-Rapson [29] and Variational Bayesian (VB) algorithms are also used instead of the IRLS algorithm. We see that a majority of the algorithms used for these models are a revised version of the EM algorithm [30] by just implementing different methods in the inner loop or gating network's maximization step. Apart from the methods discussed above, certain unsupervised learning schemes like Self Organizing Maps (SOM) and Principal Component Analysis (PCA) are used for the initial training and clustering of data in the MELE methods to obtain better accuracies.

IV.CONCLUSION

Many classification algorithms have been developed on the ME and HME architecture and have shown much better and faster results than the MLP or RBF techniques. The main advantages of these models are that unlike the MLP and other ANN algorithms, they avoid the problem of overfitting. But, in spite of these advantages, very few algorithms [16] [29] have been proposed for multiclass classifications. Not much work has been successfully done landcover classification using satellite images with these models apart from [16]. Another noticeable point stated in the paper is regarding the significance of the number of experts, since increase in the number of experts do not contribute significantly to the accuracy levels after a particular point, whereas the complexity and processing time rises exponentially. Being one of the most advanced techniques in ANN, the reason for the poor performance in this aspect is still unresolved. This leaves an open end that needs to be explored further in the case of multiclass classifications.

Numerous online algorithms using SOM [7], [31], Tree structures [20] [21] [22], and other techniques as described in [32], [33] [34] are developed to find the structure of ME and HME models that best suits the input and output data. Nguyen [35], has proposed a method where mixture of experts is inculcated with the principles of cooperativeness and co-evaluations as a separate layer. This novel method could give better performance as it could automatically decompose problems into different regions. The greatest difficulty in the ME and HME model is that a particular approach to one classification problem could be challenging and inappropriate for other classification problem. For example, [35] and [36] were found to give poor results for multi sensor land cover classification even though they were the best methods in their respective problems. A hybrid two-step approach involving MELE and MILE methods can also to be proposed and implemented. These models have also been thought of to combine different image classification algorithms (considering each algorithm as an

expert) dynamically and increase the classification accuracy. Our future work would be to come up with an online algorithm that can be used for landcover classifications, especially multisource images with a view to generalize the technique for all pattern recognition problems.

REFERENCES

[1] K. S. Fu and J. K. Mui, A survey on image segmentation, *Pattern Recognition, 13*, 1981, *3-16*.

[2] R. M. Haralick and L. G. Shapiro, Image segmentation techniques, *Computer Vision, Graphics, and Image Processing, 29*, 1985, *100-132*.

[3] N. R. Pal and S. K. Pal, A review on image segmentation techniques, *Pattern Recognition, 26*, 1993, *1277-1294*.

[4] M. Egmont-Petersen; D. de Ridder, H. Handels, Image processing with neural networks—a review, *Pattern Recognition 35*, 2002, *2279-2301*.

[5] W.E. Reddick, J.O. Glass, E.N. Cook et al., Automated segmentation and classification of multispectral magnetic resonance images of brain using arti1cial neural networks, *IEEE Transactions on Medical Imaging 16 (6)*, 1997, *911-918*.

[6] S.B. Serpico, L. Bruzzone, F. Roli, An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images, *Pattern Recognition Letters 17(13)*, 1996, *1331-1341*.

[7] J. Waldemark, An automated procedure for cluster analysis of multivariate satellite data, *International Journal of Neural Systems 8(1)*, 1997, *3-15*.

[8] D.X. Le, G.R. Thoma, H. Wechsler, Classification of binary document images into textual or non-textual data blocks using neural network models, *Machine Vision and Applications 8(5)*, 1995, *289-304*.

[9] S. Rout, S.P. Srivastava, J. Majumdar, Multi-modal image segmentation using a modified Hopfield neural network, *Pattern Recognition 31 (6)*, 1998, *743-750*.

[10] Simon Haykin, *Neural Networks-A comprehensive foundation* (Englewood Cliffs, NJ: Prentice-Hall 1999).

[11] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Application to Clustering* (New York: Marcel Dekker, 1988).

[12] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Analysis of Finite Mixture Distributions* (New York: Wiley, 1985).

[13] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, Adaptive mixtures of local experts, *Neural Computation, 3(1)*, 1991, *79–87*.

[14] M. I. Jordan and R. A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, *Neural Computation, 6(2)*, 1994, *181–214*.

[15] Ebrahimpour. R, *View-independent face recognition with mixture of experts*, PhD thesis, Institute for studies in theoretical Physics and Mathematics, 2007.

[16] Pramod. P. Nair, A multigradient algorithm using a mixture of expert architecture for land cover classification of multisensor images, *International Journal of Remote Sensing, 32(17)*, 2011, *4933-4941*.

[17] R. A. Jacobs and M. I. Jordan, Learning piecewise control strategies in a modular neural network architecture, *IEEE Transactions on Systems, Man and Cybernetics, 23*, 1993, *337–345*.

[18] S. R. Waterhouse and A. J. Robinson, Non-linear prediction of acoustic vectors using hierarchical mixtures of experts, *Advances in Neural Information Processing Systems, 7*, 1995, *835–842*.

[19] A. S. Weigend, M. Mangeas, and A. N. Srivastava, Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting, *International Journal of Neural System, 4*, 1995, *373–399*.

[20] L Breiman, J. H.Friedman, RA Olshen, C. J.Stone, *Classification and Regression Trees* (Belmont, CA: Wadsworth International Group, 1984).

[21] J. H. Friedman, Multivariate adaptive regression splines, *The Annals of Statistics, 19*, 1991, *1-141*.

[22] J. R.Quinlan, Induction of decision trees. *Machine Learning, 1*, 1986, *81-106*.

[23] Gutta S, Huang JRJ, Jonathan P, Wechsler H, Mixture of experts for classification of gender, ethnic origin and pose of human faces, *IEEE Transactions on Neural Networks, 11(4)*, 2000, *948-960*.

[24] Tang B, Heywood MI, Shepard M, Input partitioning to mixture of experts, *Proc. International joint conference on Neural Networks*, 2002.

[25] Saeed M, Ebrahimpour R, Mixture of experts: a literature survey, *Artificial Intelligence Review, 42(2)*, 2014, *275-293*.

[26] Ajit V. Rao, David Miller, Kenneth Rose, and Allen Gersho, Mixture of Experts Regression Modeling by Deterministic Annealing, *IEEE Transactions on Signal Processing, 45(11)*, 1997, *2811-2820*.

[27] V. Ramamurti and J. Ghosh, Advances in using Hierarchical Mixture of Experts for Signal Classification, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[28] S. F. Chen, R. Rosenfeld,**;** A survey of smoothing techniques for ME models, *IEEE Transactions on Speech and Audio Processing, 8 (1)*, 2000, *37-50*.

[29] K. Chen, L. Xu, H. Chi, Improved learning algorithms for mixture of experts in multiclass classification, *Neural Networks 12*, 1999, *1229-1252*.

[30] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from incomplete data via EM algorithm, *Journal of Royal Statistical Society: Series B, 39(1)*, 1977, *1-38*.

[31] Bin Tang, Malcolm I. Heywood, and Michael Shepherd. Input partitioning to mixture of experts. *Proc. IEEE International Joint Conference on Neural Networks*, 2002, *227-232*.

[32] J.J. Verbeek N. Vlassis B. Krose, Efficient Greedy Learning of Gaussian Mixture Models, *Neural Computation, 15(2)*, 2003, *469-485*.

[33] Y. Matsuyama, S. Furukawa, N. Takeda, T. Ikeda, Fast α-weighted EM learning for neural networks of module mixtures. *Proc*. IEEE *International Joint Conference on Neural Networks*, 3, 1998, 2306-2311.

[34] Saito, K.; Nakano, R., A constructive learning algorithm for an HME, *Proc*. IEEE *International Joint Conference on Neural Networks*, 3, 1998, 1268-1273.

[35] Nguyen M H, Abbass HA, Mckay RI, A novel mixture of experts model on cooperative coevolution, *Neurocomputing, 70(1-3)*, 2006, *155-163*.

[36] Ebrahimpour R, Arani SAAA, Masoudnia S, Improving combination method of NCL experts using gating network. *Neural Computing and Applications, 22(1)*, 2013, *95-101*.