Evidence-Based Mixture of MLP-Experts

Saeed Masoudnia, Mohammad Rostami, Mahdi Tabassian, Atena Sajedin, Reza Ebrahimpour

Abstract—Mixture of Experts (ME) is a modular neural network architecture for supervised learning. In this paper, we propose an evidence-based ME to deal with the classification problem. In the basic form of ME the problem space is automatically divided into several subspaces for the experts and the outputs of experts are combined by a gating network. Satisfactory performance of the basic ME depends on the diversity among experts. In conventional ME, different initialization of experts and supervision of the gating network during the learning procedure, provide the diversity. The main idea of our proposed method is to employ the Dempster-Shafer (D-S) theory of evidence to improve determination of learning parameters (which results more diverse experts) and the way of combining experts’ decisions. Experimental results with some data sets from UCI repository show that our proposed method yields better classification rates as compared to basic ME and static combining of neural network based on D-S theory.

I. INTRODUCTION

Combining classifiers is an approach to improve the performance in classification particularly for difficult problems such as those involving a considerable amount of noise, limited number of patterns, high dimensional feature sets, and highly overlapped classes. Generally, classifier selection and classifier fusion are the two types of combination [1]. While only one or a few local experts are nominated to make the decision in classifier selection, classifier fusion assumes that all classifiers are trained over the whole feature space, and thereby they are considered as complementary [2, 3]. Classifier selection has not attracted as much attention as classifier fusion. However, classifier selection is probably the better of the two strategies, if trained well [4]. One of the most popular methods of classifier selection is the mixture of experts (ME), originally proposed by [2]. The ME models the conditional probability density of the target output by mixing the outputs from a set of local experts, each of which separately derives a conditional probability density of the target output. The outputs of expert networks are combined by a gating network which is trained to select the expert(s) that is performing the best at solving the problem [5]. In the basic form of ME [2] the expert and gating networks are linear classifiers, however, for more complex classification tasks, the expert and gating networks could be of more complicated types. For instance, [6,7] proposes a face/nonface recognition model, in which they use MLPs in forming the gating and expert networks to improve the face/nonface recognition accuracy. In [8,9], authors use MLPs with one hidden layer as experts and an RBF network as the gating network, for designing compensators for intensity modulated radiation therapy [8] and for view-independent face recognition [9].

Another ensemble approach is combining classifiers based on D-S theory of evidence. Exploitation of this theory in combination of classifier outputs, has been implemented in different styles [10-13] that causes improvements in performance which some of related works is surveyed, elaborately, in section 4.

In this paper, D-S theory of evidence [14] is used for improvement of ME model. With suggestion of new performance factor based on evidence and some modification in calculation towards dividing problem space and combination of outputs of experts and gating network, new evidence-based mixture of MLP-experts model is presented. The experimental results show that a better performance can be achieved by using our proposed model in classification problem.

The paper is organized as follows. Section 2 summarizes concept of mixture of MLP experts. In Section 3 Dempster-Shafer theory of evidence is introduced. Section 4, gives a review of related works in combination method based on D-S theory. In Section 5, we introduce our new model, evidence-based ME, which is improvement in dynamic combining of neural networks using Dempster-Shafer theory of evidence. Section 6 is devoted to experimental results for...
validation of developed method over benchmark classification problems. The paper is finally concluded in Section 7.

II. Mixture of MLP experts

MLPs or Multi-Layer Perceptrons has been used successfully for solving different classification problems. However when there exists some kind of complexity, calculation of the related parameters of MLPs grow exponentially. This yields into massive computational tasks. To solve this problem, one can consider usage of the principle of divide and conquer. According to the principle of divide and conquer, a complex computational task can be solved by dividing it into some simple computational tasks and then combining the solutions of those tasks. The combination of experts is said to constitute a combination of classifiers [15].

The most famous method in the category of dynamic classifiers combining is Mixture of experts (ME). In this method the input signal is directly involved in the mechanism that integrates the output of the experts into an overall result [15]. For the first time ME was proposed in [2,16]. Their proposed model contains a population of simple linear classifiers (the experts). Output of the experts is mixed by using a gating network. Technically the experts perform supervised learning since for modeling the desired response outputs of individual experts are combined. The experts also self-organize to find a good partitioning of the input space and each expert models its own subspace, and combination of all experts model the input space well. In [17,18] this method was extended to the so-called “Hierarchical Mixture of Experts” (HME). In HME instead of single experts, mixture of experts model are used for each component. ME has been studied for a wide range of research [19-22].

Mixture of Multi-Layer Perceptrons Experts architecture: in this version of Mixture of Experts, in order to improve the performance of the expert networks, we use MLPs instead of linear networks or experts. This architecture is depicted in Fig. 1. Revision in the learning algorithm is necessary since we use MLPs in the structure of expert networks. The learning algorithm is modified by using an estimation of the posterior probability of the desired output by each expert. This way, the gating and expert networks match and this version of Mixture of Experts, in order to improve the performance of the expert networks, we use MLPs instead of linear networks or experts. This architecture is depicted in Fig. 1. The block diagram of mixture of MLP experts. It is composed of experts and gating network. The experts compete to learn the training patterns and the gating network mediates the competition

Thus the gating network computes \( O_i \), which is the output of the MLP layer of the gating network, then applies the softmax function to get:

\[
 g_i = \frac{\exp(O_i)}{\sum_{j=1}^{N} \exp(O_j)} \quad i = 1, ..., N
\]

Where \( N \) is the number of expert networks, so \( g_i \) are nonnegative and sum to 1. The final mixed output of the entire network is:

\[
 O_t = \sum_i g_i O_i \quad i = 1, ..., N
\]

The weights of MLPs are learned using the error backpropagation, BP, algorithm, in order to maximize the log likelihood of the training data given the parameters. For each expert \( i \) and the gating network, the weights are updated according to the following rules:

\[
 \Delta w_{yi} = \eta_h (y - O_t)(O_i (1 - O_i))O_{ht}^T \quad (3)
\]

\[
 \Delta w_{yi} = \eta_h (y - O_t)(O_i (1 - O_i))O_{hi}(1 - O_{hi})x_i \quad (4)
\]

\[
 \Delta w_{yi} = \eta_g (h - g)O_{yi} \quad (5)
\]

\[
 \Delta w_{yi} = \eta_g (h - g)O_{yi}(1 - O_{yi})x_i \quad (6)
\]
Where $\eta_e$ and $\eta_i$ are learning rates for the expert and the gating network, respectively. $w_e$ and $w_i$ are the weights of input to hidden and hidden to output layer, respectively, for Experts and $w_{hy}$ and $w_{oy}$ are the weights of input to hidden and hidden to output layer, respectively, for the gating network. $O_h^T$ and $O_y^T$ are the transpose of $O_h$ and $O_y$, the outputs of the hidden layer of expert and gating networks, respectively. $h_i$ is an estimation of the posterior probability that expert $i$ can generate the desired output $y$:

$$h_i = \frac{g_i \exp(-\frac{1}{2}(y-O_y)^T(y-O_y))}{\sum_i g_i \exp(-\frac{1}{2}(y-O_y)^T(y-O_y))}$$  \hspace{1cm} (7)

As pointed out by in [24], in the network’s learning process, “the expert networks “compete” for each input pattern, while the gate network rewards the winner of each competition with stronger error feedback signals. Thus, over time, the gate partitions the input space in response to the expert’s performance”.

III. DEMPSTER-SHAFER THEORY OF EVIDENCE

The history of the Dempster-Shafer theory of evidence returns to 1960, when Dempster [25] proposed his mathematical theory that later was extended by Shafer [14]. This theory is a powerful tool for representation of uncertainty and combination of evidences and generalizes traditional Bayesian reasoning with flexibility. Major concepts of this theory are briefly reviewed in the following.

Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a finite set of mutually exclusive and exhaustive hypotheses called the frame of discernment. A basic belief assignment (BBA) or mass function is a function $m: 2^\Theta \rightarrow [0, 1]$, and it satisfies the two following conditions:

$$m(\emptyset) = 0$$

$$\sum_{A \subseteq \Theta} m(A) = 1$$  \hspace{1cm} (8)

where $\emptyset$ is the empty set and a BBA that satisfies the condition $m(\emptyset) = 0$ is called normal. The subsets $A$ of $\Theta$ with nonzero masses are called the focal elements of $m$ and $m(A)$ indicates the degree of belief that is assigned to the exact set of $A$ and not any of its subsets. There are also two other definitions in the theory of evidence. They are belief and plausibility functions associated with a BBA and are defined respectively, as follow:

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B)$$

$$\text{Pl}(A) = \sum_{B \subseteq \Theta} m(B)$$  \hspace{1cm} (9)

$\text{Bel}(A)$ represents the total amount of probability that must be allocated to $A$, while $\text{Pl}(A)$ can be interpreted as the maximum amount of support that could be given to $A$, note that the three functions $m$, $\text{Bel}(A)$ and $\text{Pl}(A)$ are in one-to-one correspondence and by knowing one of them, the other two functions could be derived.

Assume two BBAs $m_1(\cdot)$ and $m_2(\cdot)$ for belief function $\text{Bel}_1(\cdot)$ and $\text{Bel}_2(\cdot)$, respectively and let they are induced by two independent sources of evidence. These pieces of evidence can be combined using Demster’s rule of combination (also called the orthogonal sum) which is defined as

$$m(C) = \sum_{\Delta \in C \subseteq \Theta} m(\Delta) \times m(\Theta - \Delta)$$

As pointed out by in [24], in the network’s learning process, “the expert networks “compete” for each input pattern, while the gate network rewards the winner of each competition with stronger error feedback signals. Thus, over time, the gate partitions the input space in response to the expert’s performance”.

IV. EXISTING EVIDENCE-BASED METHODS FOR COMBINING CLASSIFIERS

One of the early works on this topic was proposed in [26] by Mandler and Schurmann. In that method, distance measures of different classifiers were transformed to evidence. First, distance between learning datasets and some reference points was calculated and used to estimate statistical distributions of intraclass and interclass distances. Intraclass distance is a distance within a specific class label, while interclass distance is the distance between different
classes. A posteriori probability function was estimated for both distances. This indicated degree at which an input pattern belongs to a certain reference point. In next stage the clad conditional probabilities were combined into evidence ranging from 0 to 1 and this was considered as BBA of that class. Finally, by using Dempster-Shafer combination rule the BBAs of different classifiers were combined to produce final result. The benefit of this approach over traditional Bayesian method is that in Bayesian method some sort of distribution is pre-assumed while in this approach a certain statistical model is developed. In [10] it is explained that applying this method to neural networks lead to questions about a choice of the reference vectors and a distance measure. Also, approximations in estimating parameters of statistical models for intra-class and inter-class distances can lead to inaccuracy in measurement of evidence.

In [27] to perform the classification, \( K + 1 \) classes were used. In \( K + 1 \) class, the data that the classifier has no idea about which class the input comes from, are gathered. For each classifier \( e_i^n, n = 1, \ldots, N \), recognition, substitution and rejection rates \( (e_i^n, e_i^n, 1 - e_i^n - e_i^n) \) were used as a measure of BBA. Recognition rate is how accuracy of correctly classifying a test pattern, substitution rate is the misclassification rate and the rejection rate is the measure of the amount of patterns designated the class label \( K + 1 \).

Based on the above assumption in [27] an algorithm is developed to solve the problem.

In [28] Dempster-Shafer theory is used for the problem of person identification and automatic speech recognition in an applied approach. An algorithm is proposed that apprehend the fact that the performance parameter of a classifier is a good candidate to be used as evidence in favor of the classifier’s decision. The algorithm in the validation stage works on the validation database to evaluate performance parameter of the classifier, which is converted into evidence. At the testing stage, these evidences are combined using Dempster-Shafer theory to get final decision.

We use power of this theory in classifier combining, for improving structure and learning methods of ME model. In next section we provide detailed description of our proposed model.

V. EVIDENCE-BASED MIXTURE OF MLP-EXPERTS

In this section, we describe our proposed method, evidence-based mixture of MLP-experts. Dempster-Shafer theory of evidence is used for improvement of dynamic combination of neural networks classifiers. Using this theory, we have proposed a new method for calculation of \( h_i \) and combining mechanism of classifiers’ outputs in mixture of experts structure.

In this modular network every classifier is a MLP network with one hidden layer and the output \( O_i \) is produced by applying sigmoid activation function. Suppose \( \{ X^i \} \) is a subset of learning data set which belongs to \( j^{th} \) class and \( O_i^j(X) \) is the \( i^{th} \) output of neural network for \( \{ X^i \} \) as input. The average of the \( i^{th} \) output of neural network for \( j^{th} \) class can be calculated as follows:

\[
\bar{R}_i^j = \frac{\sum_{x \in \{ X^i \}} O_i^j(x)}{|\{ X^i \}|}
\]  

(16)

In which \( \bar{R}_i^j \) is the reference vector of \( i^{th} \) expert for \( j^{th} \) class. For assessing expert outputs of the system, proximity of these outputs with reference vector is calculated.

For this reason, the proximity function \( \varphi \) is defined by:

\[
s_i^j = \varphi(R_i^j, O_i^j)
\]  

(17)

It is a measure of similarity between output of \( i^{th} \) classifier and reference vector assigned to \( i^{th} \) classifier and \( j^{th} \) class.

As explained, in ME methodology considering \( h_i \) and \( g_i \) and using them for learning process, each expert is learnt on its efficiency in different subspace of input problem space. In this process, gating network is responsible for dividing problem space based on the ability of experts and weighting the outputs of each expert according to their expertise. In our proposed method, some modifications are made to improve learning parameters and method of combining outputs of experts.

\( h_i \) is defined as a measure for the ability of experts and it performs rivalry mechanism in learning process of ME. If we use a measure, proportional with the ability of classifiers, which applies more diversity on its basic classifiers, we can hope that the final combination be more efficient [29]. For this purpose, we have used this parameter as \( h_i \) coefficient in creating rivalry in competitive learning process of ME. This increases the diversity between basic experts.

Another modifications is made for improving classic ME method. For evaluation of distance we have used reference vector of each expert/class \( \bar{R}_i^j \) instead of label vector of each class. In addition to desired output information this measure encompasses information related to output pattern of classifier for each class. This will yield into more efficiency in calculation of distance.

Considering these explanation and related formulas in previous section, we calculate the evidence of class \( k \) which is obtained by expert \( i, h_i \) as follows:
is calculated based on evidence measure of corresponding class, using gating output. With this difference that, instead of specifying weights for experts, a different weight is assigned to the decision of each expert based on evidence of that class.

\[ g_{ip} = \frac{O_{ip} \prod_{j \neq p} (1 - O_{ij})}{1 - O_{ip} \prod_{j \neq k} (1 - O_{ij})} \]  

(19)

, where \( N \) is the number of experts.

To calculate final decision, for each input \( x \) per each class \( k \), \( e_k(x) \) (which implies evidence “belonging \( x \) to the class \( k \)”) is calculated by every expert and the gating weight each evidence of expert/class using this weighted combination:

\[ evidence_k(x) = \prod_{i=1}^{N} e_{ki}(O_i) \]  

(20)

Finally, the input data \( x \) is assigned to the class \( C \), which has the most evidence:

\[ C = \max(evidence_k(x)) \]  

(21)

In next section, we show the effectiveness of our method using some experimental results.

VI. EXPERIMENTAL RESULTS

Five UCI data sets [30] are used in the experiments. Information of these data sets are shown in Table I.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>SUMMARY OF DATA SETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set</td>
<td>Size</td>
</tr>
<tr>
<td>Sonar</td>
<td>208</td>
</tr>
<tr>
<td>Breast Cancer Wisconsin</td>
<td>569</td>
</tr>
<tr>
<td>Pima Indian Diabetes</td>
<td>768</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
</tr>
<tr>
<td>Vehicle</td>
<td>946</td>
</tr>
</tbody>
</table>

We have compared our proposed model versus static combining of neural networks based on D-S theory and basic ME. For each of these ensemble models three identical MLP networks, with different initial random weights, are used as experts and a gating network. The MLPs and the gating network consist of 30 and 20 sigmoid neurons in the hidden layer, respectively.

The classification performance is measured using k-fold cross validation technique, with \( k = 5 \). The whole data sets are partitioned into 5 equally-sized subsets and each subset is used as a test set for classification model that is trained on the remaining subsets. Three ensemble models, is trained via back propagation algorithm in 100 epochs and the learning rate values of experts and gating networks, \( \eta_e \) and \( \eta_g \) were considered 0.2 and 0.1, respectively.

The results are summarized in Table II. In order to show the ability of utilized combining methods versus stand alone MLP in increase of classification rate, single MLP is used in similar experiments which their results is reported in Table III. These results are the average of five times testing the corresponding models, through cross validation procedure.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>The Classification Rates of Different Ensemble Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>ME</td>
</tr>
<tr>
<td></td>
<td>Best</td>
</tr>
<tr>
<td>Sonar</td>
<td>97.60</td>
</tr>
<tr>
<td>Breast</td>
<td>98.07</td>
</tr>
<tr>
<td>Pima</td>
<td>78.26</td>
</tr>
<tr>
<td>Glass</td>
<td>65.42</td>
</tr>
<tr>
<td>Vehicle</td>
<td>81.56</td>
</tr>
</tbody>
</table>

The results confirm the better performance of our proposed model.
TABLE III
The Classification Rates of Single MLP

<table>
<thead>
<tr>
<th>Data set</th>
<th>Performance (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single MLP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Best</td>
<td>Average</td>
</tr>
<tr>
<td>Sonar</td>
<td>94.62</td>
<td>93.65 (0.1E-1)</td>
</tr>
<tr>
<td>Breast</td>
<td>97.60</td>
<td>94.45 (0.05E-1)</td>
</tr>
<tr>
<td>Pima</td>
<td>70.96</td>
<td>68.13 (0.2E-1)</td>
</tr>
<tr>
<td>Glass</td>
<td>61.68</td>
<td>60.56 (0.1E-1)</td>
</tr>
<tr>
<td>Vehicle</td>
<td>77.42</td>
<td>76.6 (0.06E-1)</td>
</tr>
</tbody>
</table>

TABLE IV
The Analysis of Outputs of Evidence-Based Gating Network in Classification of Sonar Data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$g_1$</td>
<td>$g_2$</td>
</tr>
<tr>
<td>ME</td>
<td>0.333</td>
<td>0.337</td>
</tr>
<tr>
<td>Ev. based ME</td>
<td>0.345</td>
<td>0.329</td>
</tr>
</tbody>
</table>

STD, presents standard deviation of $g_i$s in each model/class.

Each $g_i$ is (normalized average of) evidence of hypothesis that expert $i$ can generated the desired output. It is calculated by multiplication of $g_0$ of all outputs of gating network.

In the Table IV, larger standard deviation of $g_i$ in comparison with basic ME characterizes more diversity in evidence-based ME. The advantage of using D-S theory, is its power in diversifying expert networks, consequently we reach to a better performance than basic ME. Meanwhile in our proposed method, the problem space is automatically divided into subspaces which yield in better performance compared to combining of neural networks based on D-S theory.

The performance curves for each of three models, through learning process of classification of Breast Cancer Wisconsin data set, are shown in Fig. 2 which asserts supremacy of our proposed model.

VII. CONCLUSION

A modified version of mixture of experts in the framework of D-S theory was proposed. We employed D-S theory to improve determination of learning parameters and combining of the decisions of experts. Experimental results on five data sets from UCI Repository demonstrated the advantage of our proposed method in comparison with basic ME and static combining of neural networks based on D-S theory. The efficiency of our method over the static combining of neural networks can be referred to divide and conquer principle used in the ME structure. Regarding the classification results of the proposed method and those of basic ME, better performance of our classification scheme can be interpreted by consideration of generating more diverse experts using D-S theory.

In addition, by using combining methods, we get significantly better performance than stand alone MLP, which shows that classifier combining can generate more accurate classification than each of constituent classifiers[4].

REFERENCES