Mixture of Random Prototype-based Local Experts

Nima Hatami

Dept. of Elec. & Elec. Eng.

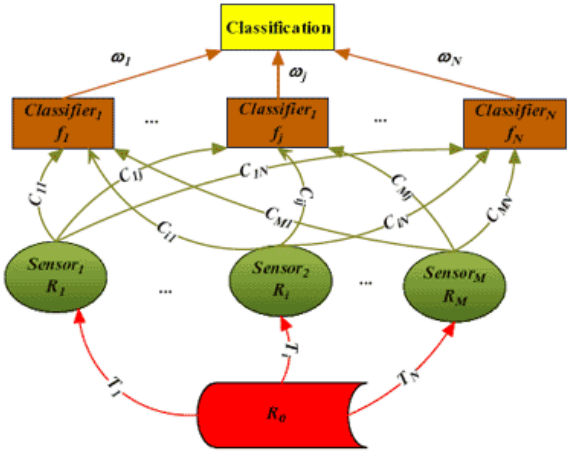
Univ. of Cagliari, Italy

Agenda

- Classifier ensembles
- Mixture of Experts (ME) model
- Mixture of Random Prototype-based Experts
- Experimental results
- Conclusions & Future works

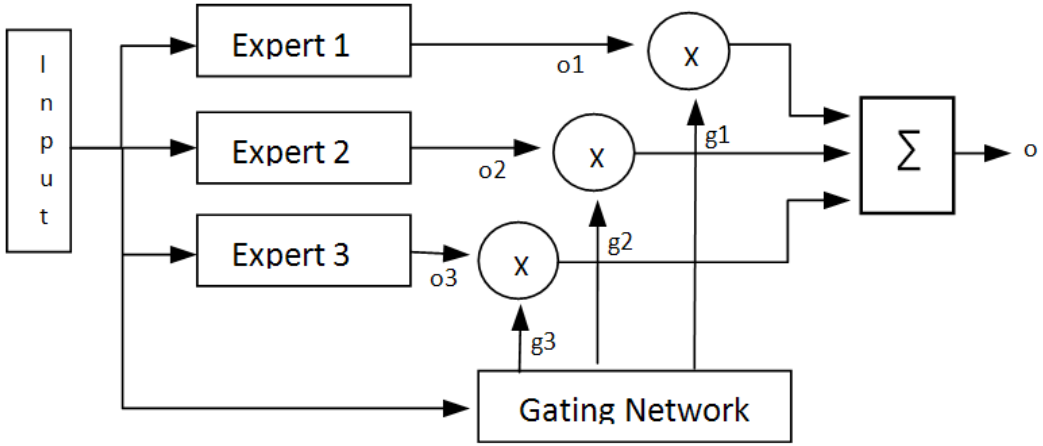
Classifier ensembles

- also Known as classifier fusion, combining classifiers, MCS
- Most real-world PR problems are too complicated for a single classifier to solve
- Divide-and-conquer has proved to be efficient in many of these complex situations
- combination of classifiers which have complementary properties



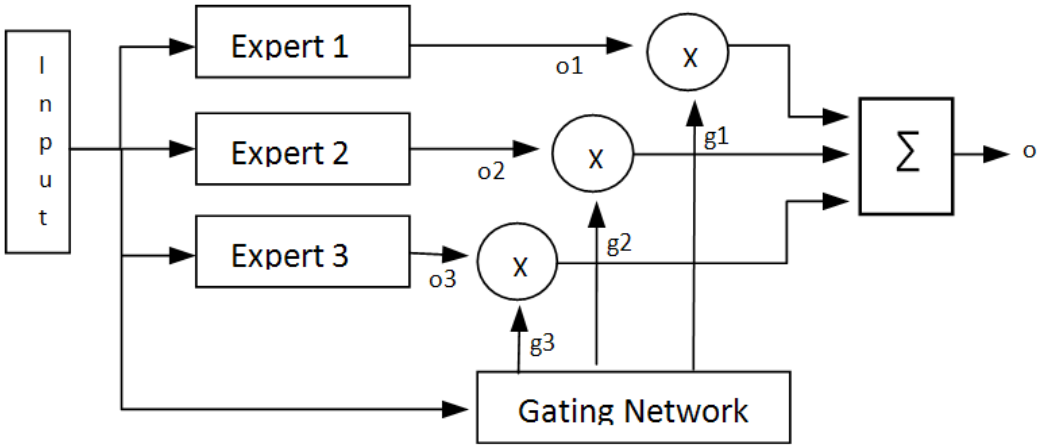
Mixture of Experts

- Jacobs et al. have proposed the ME based on the divide-and-conquer strategy
- one of the most popular ensemble methods used in PR and ML
- a set of expert networks is trained together with a gate network



Mixture of Experts

- **stochastically** partitions the input space of the problem into a number of subspaces
- experts becoming specialized on each subspace
- uses the gating network to manage this process



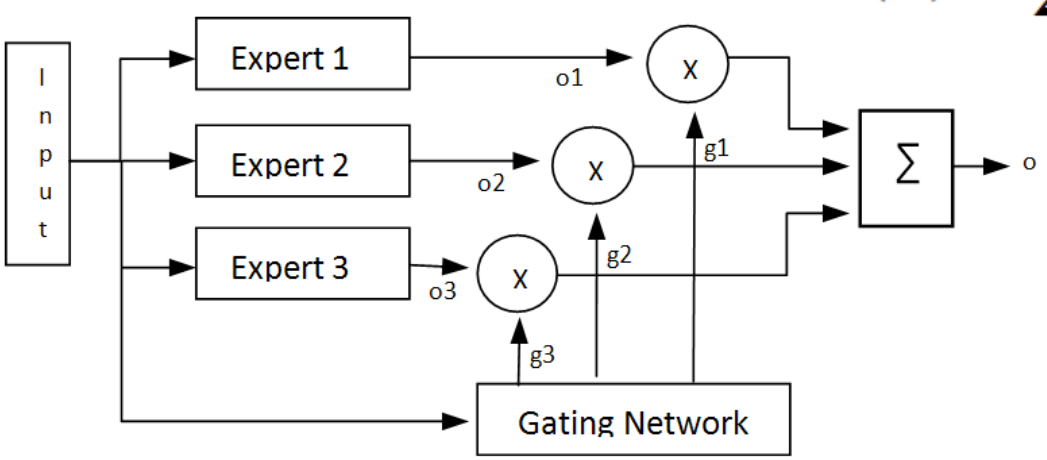
Mixture of Experts

- *i*th expert
- Gating network
- Final output

$$o_i(x, W_i) = f(W_i x)$$

$$g_i(x, V_i) = \frac{\exp(o_{g_i})}{\sum_{i=1}^{N} o_{g_j}} \quad i = 1, \ldots, N$$

$$o(x) = \sum_i g_i(x, V_i) o_i(x, W_i)$$



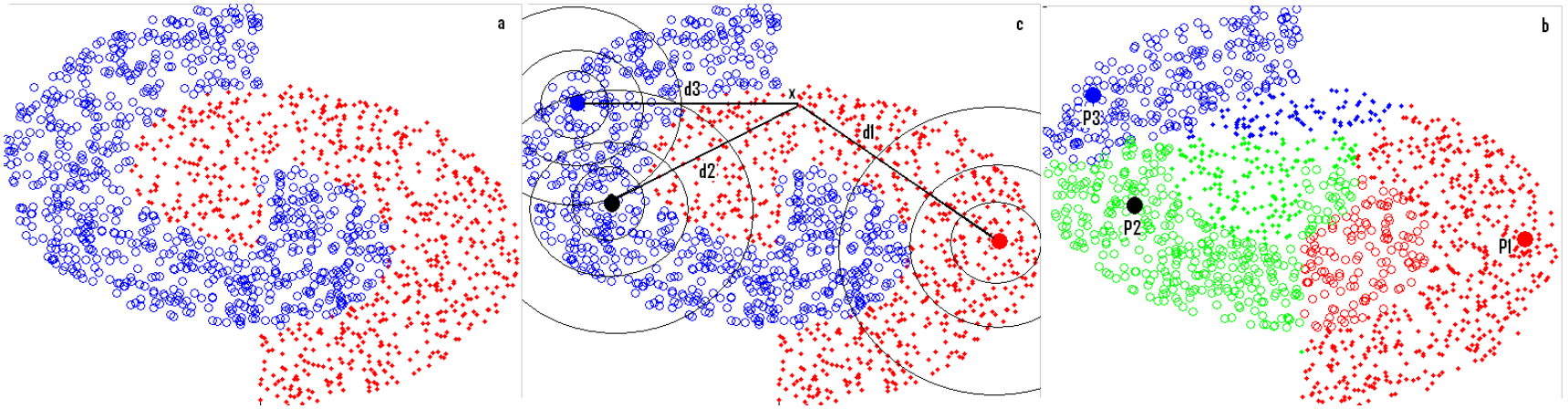
Mixture of Experts model

- **Why does ME succeed?**

1. Encourages **diversity** between the single experts by automatically localizing them in different regions of the input space
2. achieves **good combination weights** of the ensemble members by training the gate, which computes the dynamic weights together with the experts

Random Prototype-based Data Splitting

- selects some prototype points from the input space
- partitions this space according to nearest distance from these prototypes
- two different partitioning methods, i.e. disjoint and overlapping



Mixture of Random Prototype-based Experts

- Earlier works on the ME apply methods such as preprocessing to partition the input space or transform the input space into simpler and more separable spaces
- a modified version of the ME algorithm
- partitions the original problem into **centralized** regions
- uses a simple distance-based gating function to specialize the expert networks (**training step**)
- Contribution of each expert is according to the distance between the input and a prototype embedded by the expert (**testing step**)

Mixture of Random Prototype-based Experts

Mixture of Random Prototype – based Experts Algorithm

INITIALIZING:

- $P = \{p_i \in LS \,|\, i = 1,2,\ldots,N\}$; LS = Learning Set, TS = Testting Set
- $\psi = \{\varepsilon_i \,|\, i = 1,2,\ldots,N\}$
- strategy = {static, dynamic}
- $E = \{\eta_j \in (0,1) \,|\, j = 1,2,\ldots,N\}$ such that:

$$\eta_k \leq \eta_{k+1}; k = 1,2,\ldots,N-1 \text{ and } |E| = \sum_j \eta_j = 1$$

TRAINING:

For each $x \in LS$ Do:

- $D(x) = \{d_i(x) \,|\, i = 1,2,\ldots,N\}$ where

  $d_i(x) = \|x - p_i\|$ and $\|\cdot\|$ is any distance metric (e.g. Euclidean)
- $H(x) = \{h_i(x) \,|\, i = 1,2,\ldots,N\}$ where

  $h_i(x)$ represents the expected capacity of ε_i to deal with the given input x

  [strategy = static]: $h_i(x) = \eta_j$ where $J = \text{Rank}(\varepsilon_i, D(x))^*$

  [strategy=dynamic]: $h_i(x) = 1 - \frac{d_i}{|D(x)|}$ where $|D(x)| = \sum_j d_j(x)$
- update each expert ε_i $(i = 1,2,\ldots,N)$ according to the standard learning rule for ME

TESTING:

Given an $x \in TS$ Do:

- $D(x) = \{d_i(x) \,|\, i = 1,2,\ldots,N\}$
- $G(x) = \{g_i(x) \,|\, i = 1,2,\ldots,N\}$ where

  [strategy=static]: $g_i(x) = \eta_j$ where $j = \text{Rank}(\varepsilon_i, D(x))^*$

  [strategy=dynamic]: $g_i(x) = 1 - \frac{d_i}{|D(x)|}$ where $|D(x)| = \sum_j d_j(x)$
- calculate the overall output:

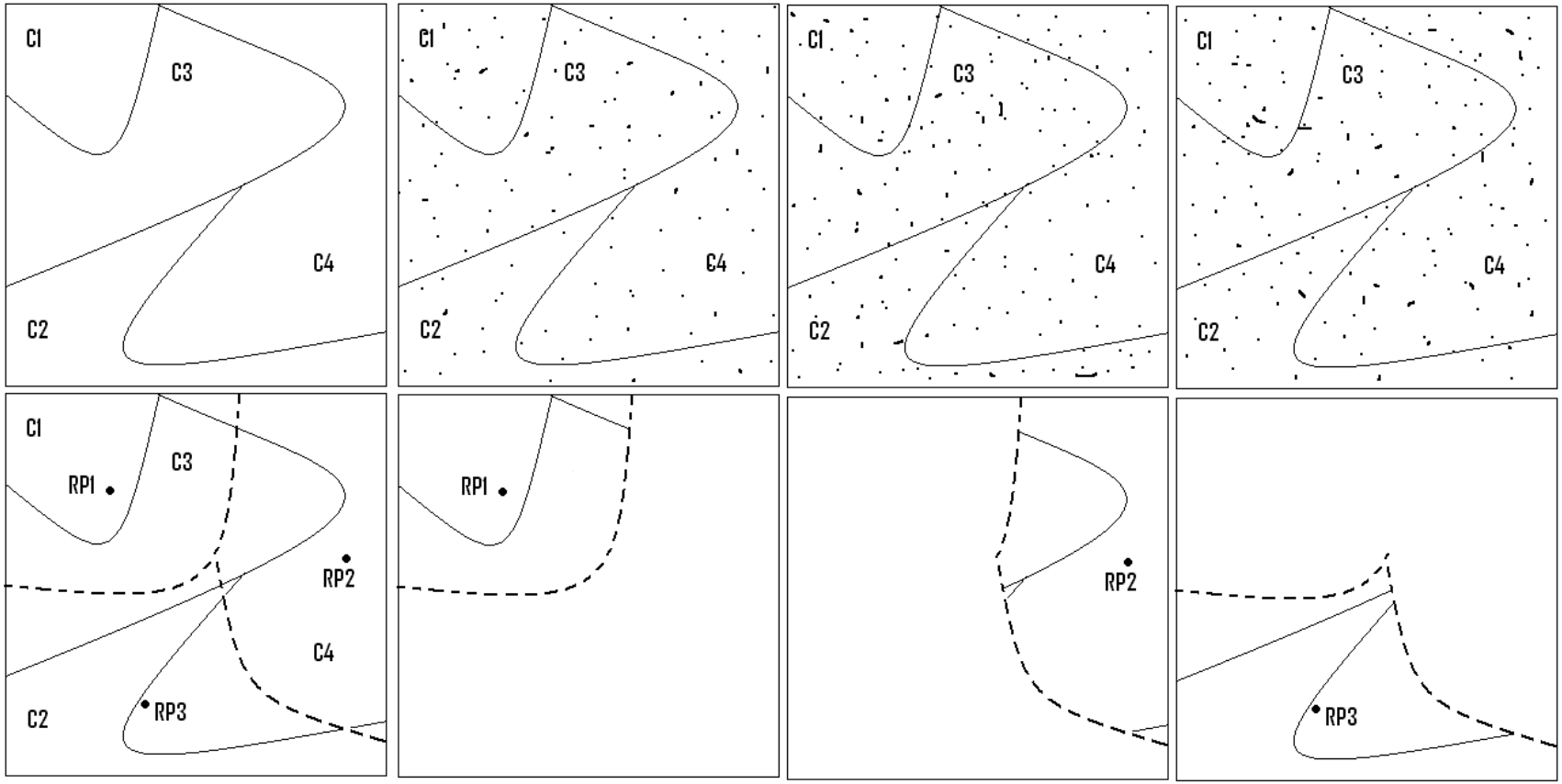
$$o_j(x) = \sum_{i=1}^{N} g_i(x).o(x, W_i)$$

- select the class label c_k such that

$$k = \arg\max_j (o_j(x))$$

* $j = \text{Rank}(\varepsilon_i, D(x))$ returns the rank of expert ε_i (i.e. a number in $[1, N]$) according to the distance $D(x)$ evaluated on the input x (the lovest the distance, the highest the ranking)

ME vs. MRPE on a toy problem



The ME vs. MRPE

- Resulting misclassifications in the standard ME derive from two sources:

1. the gating network is unable to correctly estimate the probability for a given input sample
2. local experts do not learn their subtask perfectly

The ME vs. MRPE

- Improves three important aspects of the standard ME model
  1. reduces the **training time** by decreasing the number of parameters to be estimated
  2. as simple distance measures used by the gating function are more robust with respect to **errors in determining the area of expertise** of an expert, errors in the proposed ME model are mainly limited to the error made by the expert networks
  3. the area of expertise of each expert is more centralized, which makes the subproblem **easier to learn**

Experimental results

- We used some of the UCI ML datasets
- 10-fold cross-validation
- Multi-layer perceptron (MLP)
- For *N*, number of partitions (experts), we varied it from 2 to 10

Table 1. The main characteristics of the selected UCI datasets

Problem	# Train	# Test	# Attributes	# Classes
Iris	150	-	4	3
Satimage	4435	2000	36	6
Pendigits	7494	3498	16	10
Letter	20000	-	16	26
Vowel	990	-	11	11
Segment	210	2100	19	7
Glass	214	-	9	7
Yeast	1484	-	8	10

Experimental results

Table 2. The mean and standard deviation of accuracy of the ME vs. the proposed mixture of random prototype-based experts on the selected UCI datasets (in percentage)

	Iris	Sat.	Pen.	Lett	Vow.	Seg.	Gla.	Yeast
Standard ME	87.7± 0.61	88.7± 1.05	88.0± 0.43	70.9± 0.93	61.1± 1.05	79.2± 0.95	72.3± 1.65	49.3± 2.01
Disjoint partition	88.2± 0.45	90.1± 0.83	89.0± 0.44	72.0± 0.80	62.9± 1.11	81.9± 0.79	74.8± 1.76	50.7± 1.96
Overlapping partition	88.5± 0.39	90.1± 0.79	89.2± 0.40	72.8± 0.95	63.4± 1.20	81.9± 0.83	75.5± 1.57	52.0± 1.95

Table 3. Training time of the ME vs. the proposed mixture of random prototype-based expert classifiers (seconds)

	Iris	Sat.	Pen.	Lett.	Vow.	Seg.	Gla.	Yeast
Standard ME	50	232	351	324	59	49	30	41
Proposed method	28	158	221	258	39	32	21	29

Conclusions

- A modified version of the popular ME algorithm is presented
- specializes expert networks on centralized regions of input space instead of nested and stochastic regions
- Using simple distance-based gating thus reduces the network complexity and the training time
- Improves overall classification accuracy

Future works

- defining a procedure for automatically determining the **number of optimal experts** for each problem without resorting to complex preprocessing
- investigate application of the proposed method to **HME** structure
- Adaptation of this method to **simple distance-based classifiers** instead of NNs
- **heuristics** able to help in the process of partitioning the input space instead of using RP

Thanks 4 ur attention! :-)

Question? :-?

Contact:

nima.hatami@diee.unica.it