Group-wise Feature Selection for Supervised Learning

Qi Xiao [1] Hebi Li[2] Jin Tian [2] Zhengdao Wang [1]

[1]ECpE at Iowa State University [2]CS at Iowa State University

Paper ID: 5752

Group-wise Feature Selection (GroupFS)

A new problem setting between global feature selection and instance-wise feature selection

Formally, find a mapping for $S : \mathcal{X} \to F$ and a set of feature selectors $F = \{s^1, \ldots, s^K, s^k \in \{0,1\}^d\}$ such that for almost all $x \in \mathcal{X}$, we have

$$P(y|\mathbf{x} \odot \mathbf{S}(\mathbf{x})) = P(y|\mathbf{x}) \tag{1}$$

Assumption space:

- Global FS, $|F| = 1$; simple and interpretable but not expressive
- Instance-wise FS, $|F| = 2^d$; expressive but lack of global interpretability
- GroupFS, $|F| = K$; both expressive and interpretable

Related Works

Feature selection using regularized Mixture of Experts (MoE)

- l_2-penalized maximum-likelihood estimator to select features in MoE [3]
- EM algorithm with coordinate ascent to generate sparse solutions [1]

Limitations:

- Individual regularizer for each predictor, require a complex EM training procedure
- Both papers focus only on linear experts

Proposed method I - INVASE + Clustering

We propose a two-step method for GroupFS

1. Train an instance-wise feature selector. Each data sample has an individual feature selector.
2. Apply the K-means clustering to all the feature selectors.

Group-wise feature selector: the assigned cluster center.

Proposed method II - GroupFS with Mixture of Experts Selector

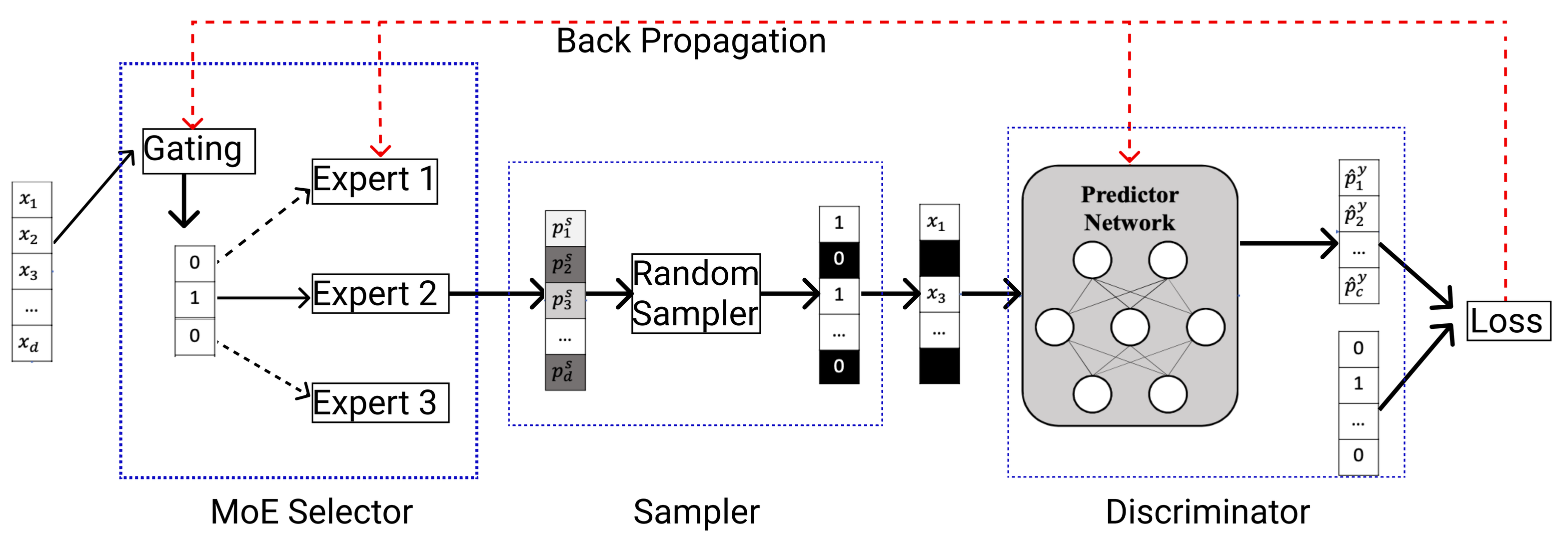

Figure 1. GroupFS-MoE Model Architecture

GroupFS-MoE: Feature Importance Score

- $s \in \{0,1\}^d$ is discrete, cannot back-propagate gradient
- Approximation: feature importance score $w \in [0,1]^d$
- Re-parametrization: $w = \text{sigmoid}(v) = \frac{1}{1+\exp(-v)}$
- Feature selector s follows Bernoulli distribution with pdf

$$\pi(s; w) = \prod_{i=1}^{d} w_i^{s_i}(1-w_i)^{(1-s_i)} \tag{2}$$

GroupFS-MoE: Mixture of Experts Selector

Mixture of K feature selectors with feature importance scores $\{w^1, \ldots, w^K\}$

$$\pi(s|x; \theta, w^1, \ldots, w^K) = \sum_{k=1}^{K} g_k(x;\theta)\pi_k(s; w^k), \tag{3}$$

$$\sum_{k=1}^{K} g_k(x;\theta) = 1, g_k(x;\theta) \in \{0,1\}. \tag{4}$$

GroupFS-MoE: Gumbel-Softmax Re-parametrization

$g(x;\theta)$ is one-hot →no gradient
Solution: Gumbel-softmax Re-parametrization

$$g_k(x;\theta) = \frac{\exp(\tau^{-1}(\log(o_k)+b_k))}{\sum_{j=1}^{K}\exp(\tau^{-1}(\log(o_j)+b_j))}. \tag{5}$$

where $b_1, \ldots, b_K \sim$ Gumbel(0,1), $o_1, \ldots, o_k$ are the original outputs of g.

Experiments: Synthetic Datasets

$$P(y=1|\mathbf{x}) = \frac{1}{1+d_i(\mathbf{x})}, \mathbf{x} \in \mathbb{R}^{11} \tag{6}$$

Syn1:

$$d_1(x) = \begin{cases} \exp(x_1 x_2), & x_{11} < 0 \\ \exp(\sum_{i=3}^{6} x_i^2 - 4), & \text{otherwise} \end{cases} \tag{7}$$

Syn2:

$$d_2(x) = \begin{cases} \exp(x_1 x_2), & x_{11} < 0 \\ \exp(-10\sin 2x_7 + 2\|x_8\| + x_9 + \exp(-x_{10})), & \text{otherwise} \end{cases} \tag{8}$$

Syn3:

$$d_3(x) = \begin{cases} \exp(\sum_{i=3}^{6} x_i^2 - 4), & x_{11} < 0 \\ \exp(-10\sin 2x_7 + 2\|x_8\| + x_9 + \exp(-x_{10})), & \text{otherwise} \end{cases} \tag{9}$$

Evaluation metric: Mean Squared Error (MSE), Accuracy (Acc) and Normalized Mutual Information (NMI):

$$\text{NMI}(C_1; C_2) = \frac{2\text{I}(C_1; C_2)}{\text{H}(C_1) + \text{H}(C_2)}. \tag{10}$$

Synthetic Results

Table 1. Learned GroupFS Feature Selectors for Syn1,Syn2,Syn3

	Syn1			Syn2			Syn3		
Experts	E1	E2	g	E1	E2	g	E1	E2	g
#samples	1617	1717		1621	1713		1621	1713	
x1	1	-	-	1	-	-	.02	.01	-
x2	1	-	-	1	-	-		.02	-
x3	-	1	-	.02	.01	-	1	-	-
x4	-	1	-	.01	.01	-	1	-	-
x5	-	1	-	.01	.01	-	1	-	-
x6	-	1	-	.02	.01	.01	1	-	-
x7	.02	-	-	-	1	-	-	1	-
x8	.03	.01	-	-	.99	-	-	1	-
x9	.03	-	-	-	1	-	-	1	-
x10	.03	-	-	-	1	-	-	1	-
x11	.17	1	.21	.02	.05	.27	.99	.99	.33

Table 2. Evaluation of proposed methods on synthetic datasets

	Syn1			Syn2			Syn3		
	NMI	MSE	Acc	NMI	MSE	Acc	NMI	MSE	Acc
INVASE+KM	.828	.189	.703	.904	.178	.715	.925	.136	.810
GroupFS-MoE	**.911**	.182	.710	**.921**	.177	.715	**.960**	.131	.811

Experiments - Real Datasets

- Boston housing: d = 13, n = 506
- Baseball salary: d = 16, n = 337
- Compare with feature selection in MoE: Khalili [3] and lasso+l_2 [1]
- Assumption: two groups of feature selector

Table 3. Discriminator's Mean Square Error (MSE) comparison with Regularized MoE

	Training				Testing	
	Khalili[3]	lasso+l_2[1]	GroupFS	INV+KM	GroupFS	INV+KM
Boston	.2044	.1989	.0879	.0853	.1863	.1846
Baseball	1.1858	.2821	.2371	.2480	.3056	.3417

INV+KM is short for INVASE+KMeans. GroupFS is short for GroupFS-MoE.

References

[1] Faicel Chamroukhi and Bao-Tuyen Huynh. Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *Journal de la société française de statistique*, 160(1):57–85, 2019.

[2] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

[3] Abbas Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4):519–539, 2010.

[4] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[5] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.