Feature Selection for Multi-labeled Variables via Dependency Maximization

Salimeh Yasaei Sekeh and Alfred O. Hero

ECE Department, University of Michigan, Ann Arbor

Contributions

- Complementarity selection criterion for dependent features
- Based on geometric mutual information (GMI)
- GMI estimated by single MST over all classes
- Computation and accuracy improvements demonstrated

Feature Selection

- A feature vector $\mathbf{X} = \{X^{(1)}, X^{(2)}, \dots, X^{(d)}\}$
- A multiclass label $Y \in \{c_1, c_2, \dots, c_m\}$
- Approaches to feature selection
  - PCA, PLS, SIR, SPARCS: linear response model
  - SVM, CI, LASSSO: categorical response model
  - Infromation divergence: empirical estimator [1]
  - Mutual information: empirical estimator (!)

Geometric Mutual Information

Conditional joint distributions:

$$\begin{aligned} f_{ij|Y} &= f(x^{(i)}, x^{(j)}|y) && \text{(dep.)} \\ \pi_{ij|Y} &= f(x^{(i)}|y) f(x^{(j)}|y) && \text{(indep.)} \end{aligned} \qquad (1)$$

Marginal joint distributions:

$$\begin{aligned} f_{ij} &= \sum_y f(x^{(i)}, x^{(j)}|y) p(y) \\ \pi_{ij} &:= \pi(X^{(i)}, X^{(j)}) = \sum_y p_y f(x^{(i)}|y) f(x^{(j)}|y) \end{aligned} \qquad (2)$$

Henze-penrose divergence between f, g:

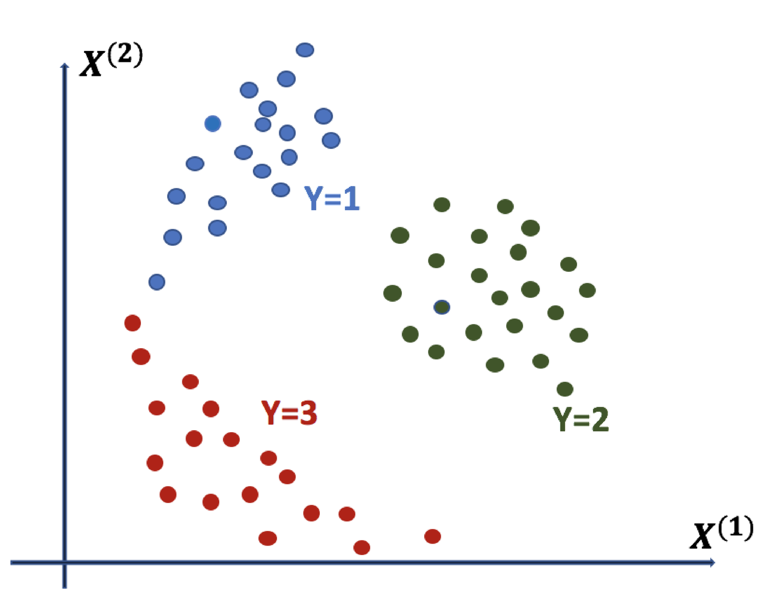
$$D(f;g) = 1 - 2\int \frac{fg}{f+g}\, \mathrm{d}\mu \qquad (3)$$

Conditional GMI:

$$I(X^{(i)}; X^{(j)}|Y) = \mathbb{E}\left[D(f_{ij|Y}; \pi_{ij|Y})\right] \qquad (4)$$

Marginal GMI:

$$I(X^{(i)}; X^{(j)}) = D(f_{ij}; \pi_{ij}) \qquad (5)$$



Conditional vs Marginal GMI

The conditional GMI is bounded by marginal GMI:

Theorem 1 *Consider conditional probability densities $f(x^{(i)}, x^{(j)}|y)$, $f(x^{(i)}|y)$, and $f(x^{(j)}|y)$ with priors p_y $y = 1, 2, \dots, m$. Then*

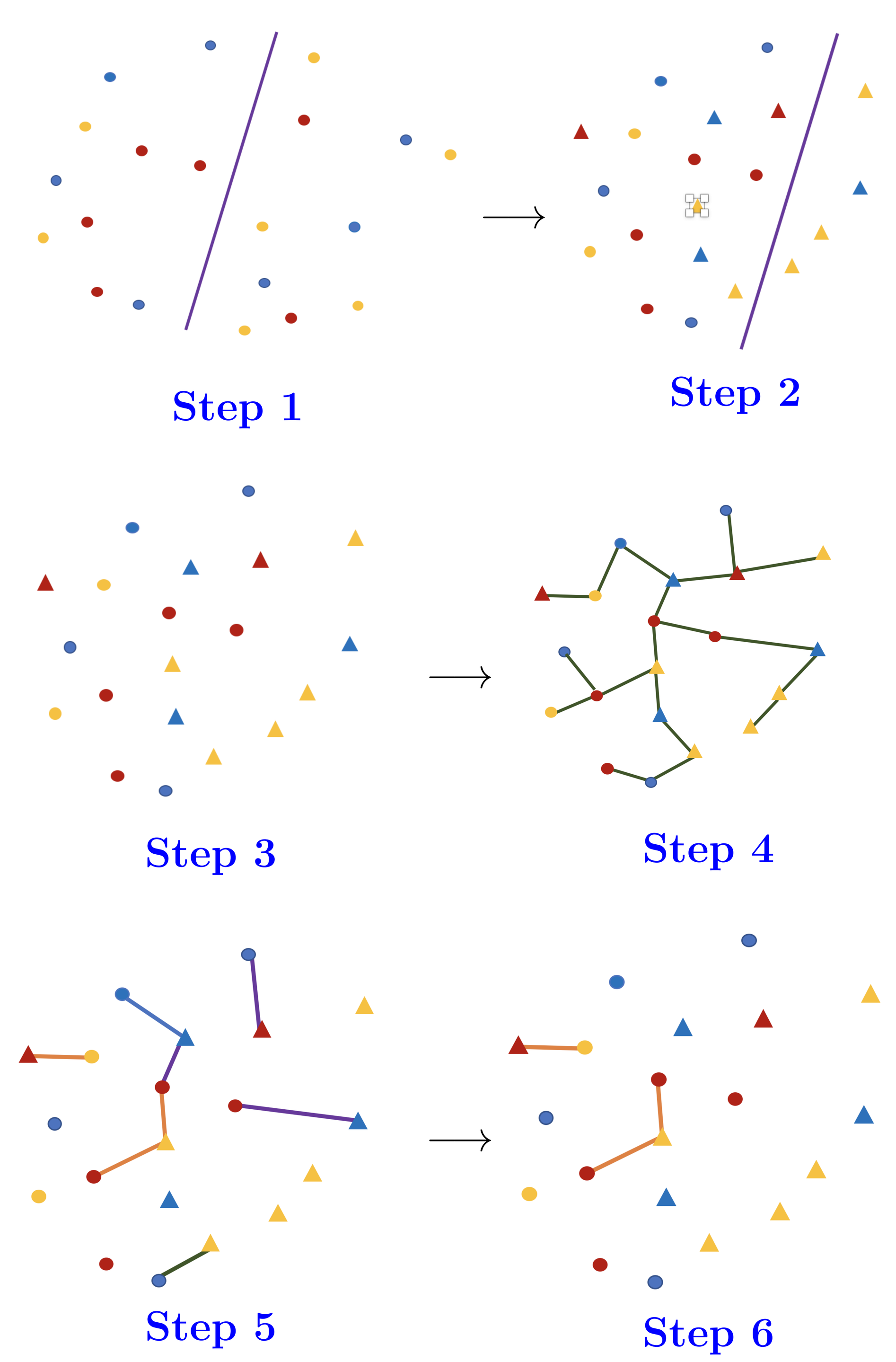
$$I(X^{(i)}; X^{(j)}|Y) \geq I(X^{(i)}; X^{(j)}) \qquad (6)$$

Complementarity feature selection criterion:

$$\rho(X^{(i)}) = \sum_{j \neq i} I(X^{(i)}, X^{(j)})$$

Multiclass GMI Estimator

Given points from features $(X^{(i)}, X^{(j)})$ with three labels:



- **Step 1**: Split data in two equal sets
- **Step 2**: Shuffle points in the second set (triangles)
- **Step 3**: Merge all points in one single set
- **Step 4**: Construct minimal spanning tree (MST) over all
- **Step 5**: Remove non-dichotomous edges
- **Step 6**: Count dichotomous edges connecting each pair of distinct labels

Estimation of Marginal GMI

Denote

- n_i: # of points in the first set (circle) with label y_i
- n_j: # of points in the second set (triangle) with label y_j

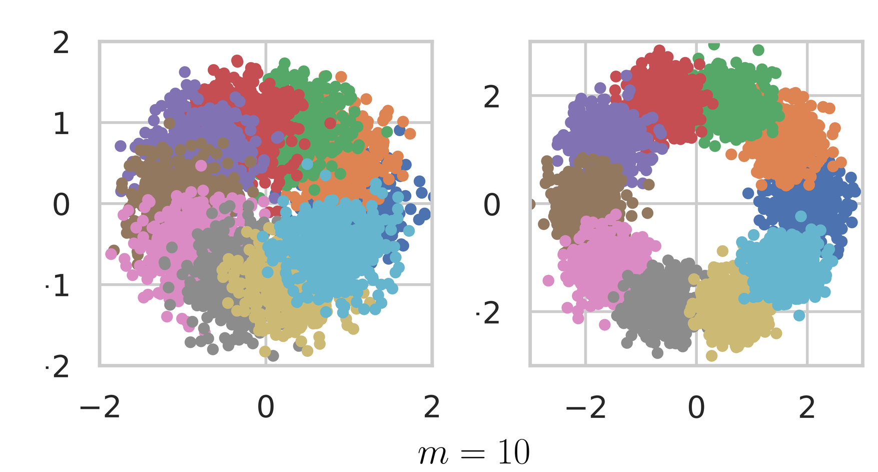
Theorem 2 *For $y_i \neq y_j$, $y_i, y_j \in \{1, \dots, m\}$, as $n_j \to \infty$, $n_i \to \infty$, $n \to \infty$ such that $n_j/n \to p_{y_j}$, $n_i/n \to p_{y_i}$, $(n = n_i + n_j)$ we have*

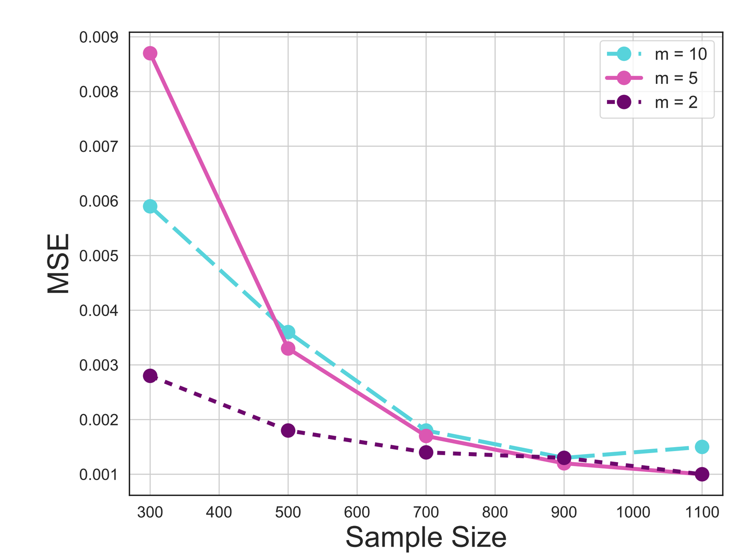
$$\left(\frac{n}{2\, n_j\, n_i}\right) \mathfrak{R}_{z_j, y_i} \longrightarrow D(f_{ij}; \pi_{ij}) \quad (a.s.) \qquad (7)$$

Numerical Experiments

- Samples drawn from one of the cases: $\mathcal{N}(\mu_i, 0.1I)$.

$$\mu_i = \left[\mu \cos\left(2\pi \frac{i}{m}\right), \mu \sin\left(2\pi \frac{i}{m}\right)\right], \quad m = 2, 5, 10$$

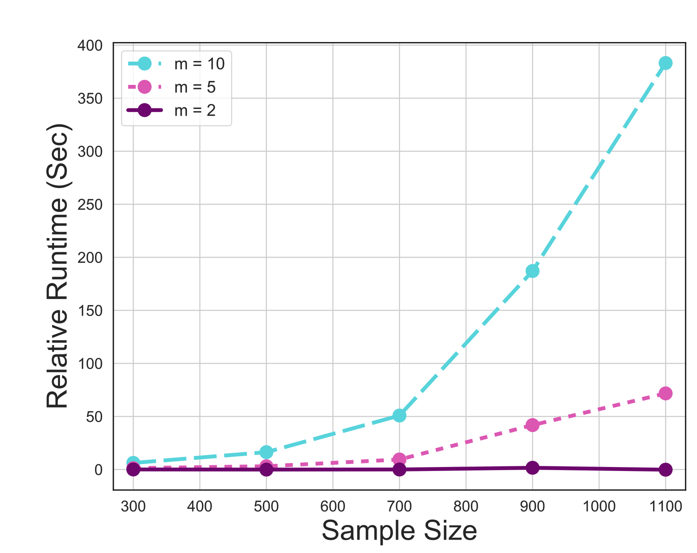




Observe:

- MSE decreases rapidly in sample size.
- As the number m of labels grow the MSE increases.

Computational complexity comparison:



Note: For large number of classes proposed method has faster runtime than Berisha et al's method (Dp algorithm) [1].

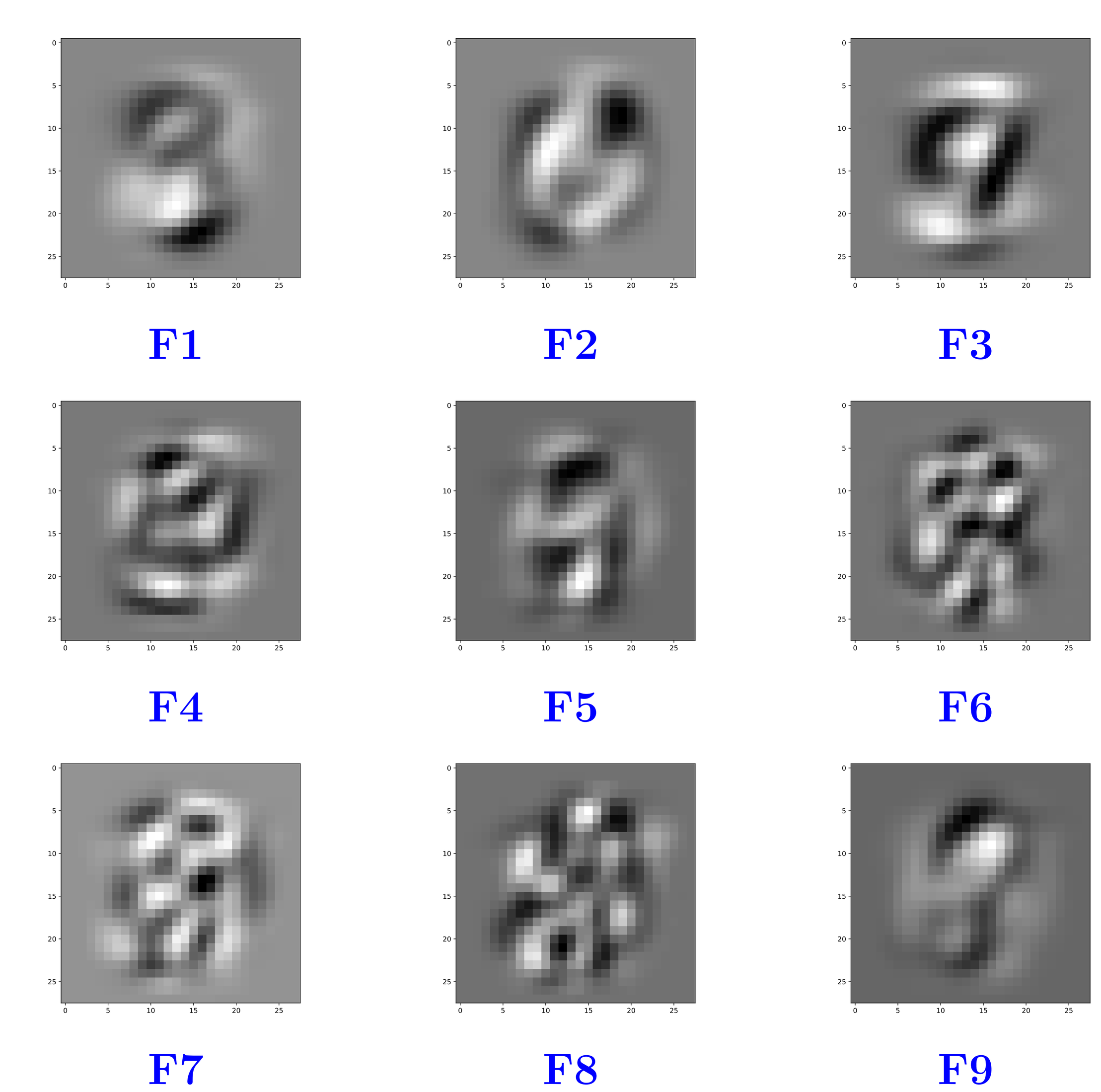
Experiments on MNIST Dataset

- 70,000, 28×28 grey-scale images of hand-written digits 0 - 9.
- Training set = 60,000 and test set = 10,000.

Number of Features	Algorithm	Number of Training Sample 100	300	500
10	GMI	**61.48**	**61.47**	**60.43**
	Dp	57.31	51.57	55.53
	LSVC	20.00	5.99	8.40
	ETC	10.69	6.00	7.09
15	GMI	**70.01**	**69.94**	66.48
	Dp	64.86	69.90	**71.71**
	LSVC	22.26	9.86	10.51
	ETC	22.26	9.84	10.51
20	GMI	73.99	73.94	72.27
	Dp	**78.95**	**77.83**	**76.77**
	LSVC	22.4	9.92	13.42
	ETC	24.67	9.93	12.77

Average classification accuracies of top features selected by GMI, Linear Support Vector Classification (LSVC), Extra-Tree-Classifier (ETC), and pairwise Dp statistic of Berisha et al [1].

- Accuracy of GMI feature selection outperforms others.
- Computational complexity of GMI is lower than pairwise Dp.



Acknowledgments

This work was partially supported by ARO under grant W911NF-15-1-0479 and USAF under grant FA8650-15-D-1845.

References

[1] V. Berisha, A. Wisler, A. Hero, and A. Spanias, Empirically estimable classification bounds based on a nonparametric divergence measure, IEEE Trans. on Signal Process. vol. 64, no. 3, pp. 580-591, 2016.

[2] S. Yasaei Sekeh and A. O. Hero, Feature Selection for Multi-labeled Variables via Dependency Maximization, arXiv:1902.03544.