A Fast Multivariate Symmetrical Uncertainty based Heuristic for High Dimensional Feature Selection

Miguel García Torres, Federico Divina, Francisco A. Gómez Vela, José L. Vázquez Noguera
-PINV18-1129-

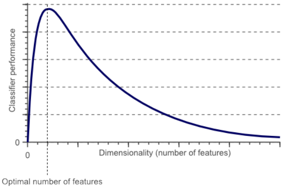
Universidad Pablo de Olavide
Universidad Americana
Data Science & Big Data Research Lab

5-7 May 2021 - Entropy 2021: The Scientific Tool of the 21st Century

1. Feature selection
   - Feature relevance and redundancy
   - Multivariate SU based Feature Selection Heuristic
2. Computational results
   - Microarray data
   - Melanoma data
3. Conclusions

Introduction
Feature selection

#	Sepal length (cm)	Sepal width (cm)	petal length (cm)	petal width (cm)	type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.4	1.3	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
...					
101	6.3	5.8	7.1	2.5	*Iris virginica*
102	6.4	2.7	5.1	1.9	*Iris virginica*
103	6.9	3.0	5.9	2.1	*Iris virginica*



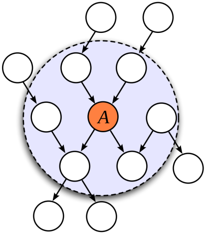
$\Rightarrow$ The objective of feature selection is to find the subset of features $S \in X$ with which $\mathcal{C}$ achieves the lowest error rate.

A feature is considered **irrelevant** if it contains no information about the class.

Markov blanket

Given a feature X_i, $M_i \subset \mathcal{X}$ ($X_i \notin M_i$) is said to be a Markov blanket for X_i iff

$$P(\mathcal{X} - M_i - \{X_i\}, \mathcal{Y} | X_i, M_i) = P(\mathcal{X} - M_i - \{X_i\}, \mathcal{Y} | M_i).$$



Feature redundancy

Entropybased measures

- Mutual Information of a given variable X with respect to variable Y (MI(Y;X)) measures the reduction in uncertainty about the value of X given the value of Y.

$$MI(X|Y) = H(X) - H(X|Y).$$

- MI is biased in favor of r.v. with more values $\Rightarrow$ Normalize.

$$SU(X, Y) = 2\left[\frac{MI(X|Y)}{H(X) + H(Y)}\right].$$

Feature relevance

Given the threshold δ, a feature X_i is relevant if its correlation with the class Y is $SU(X_i, Y) > \delta$.

Feature redundancy: Approximate Markov blanket

Given two features X_i and X_j ($i \neq j$) so that $SU(X_j, \mathcal{Y}) \geq SU(X_i, \mathcal{Y})$, then X_j forms an approximate Markov blanket for X_i iff $SU(X_i, X_j) \geq SU(X_i, \mathcal{Y})$.

The Multivariate SU

Total Correlation (TC) is a generalization of MI:

$$C(X_{1:n}) := \sum_{i=1}^{n} H(X_i) - H(X_{1:n}).$$

TC is the amount of information shared among the variables in the set.
MSU is define as follows:

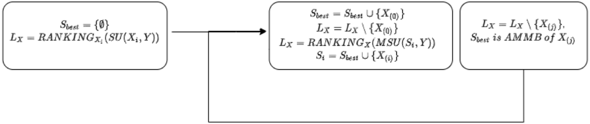
$$MSU(X_{1:n}) := \frac{n}{n-1}\left[\frac{C(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right] = \frac{n}{n-1}\left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right].$$

with $H(X_{1:n})$ the joint entropy of then random variables and $C(X_{1:n})$ defined as

MSU based Feature Selection (MSUFS) Heuristic

Approximate multivariate Markov blanket

Given two features X_i and X_j ($i \neq j$), let $S_i \subset X$ so that $X_i \in S_i$ and $X_j \notin S_i$. Then, S_i forms an approximate multivariate Markov blanket for X_j iff $SU(X_i, X_j) \geq SU(X_j, \mathcal{Y})$ or $MSU(S_i, X_j) \geq SU(X_j, \mathcal{Y})$.



Proposed heuristic workflow

Experiments

1. Microarray data. Study the performance of MSUFS on smal high-dimensional data.
2. Case study: melanoma data. Inspet the relevant features on melanoma data.

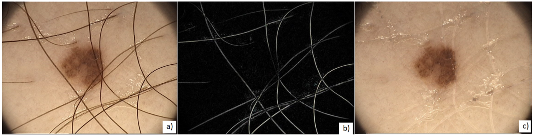
In all experiments the results achieved with MSUFS were compare with those obtained by FCBF.

dataset	id	#inst.	#feat.	labels	#inst./label
colon	cln	62	2000	normal/tumor	22/40
lymphoma	lym	77	2647	diffuse/follicular	58/19
breast	bcg	168	2905	good/poor	111/57
prostate	prt	1545	10935	tumor/not	52/50
lung	lng	181	12533	MPM/ADCA	31/150
breast	bcc	118	22215	positive/negative	75/43
breast/colon	bco	104	22283	breast/colon	62/42

id	naive Bayes			Logistic Regression		
	$\mathcal{C}$	FCBF	MSUFS	$\mathcal{C}$	FCBF	MSUFS
cln	54.62 (17.68)	75.77 (8.09)	72.82 (9.99)	83.97 (5.25)	79.10 (3.94)	77.44 (8.72)
lym	81.92 (4.93)	87.00 (4.74)	80.42 (10.69)	94.83 (5.53)	88.17 (13.65)	80.42 (10.69)
bcg	70.20 (13.75)	71.44 (11.27)	73.74 (12.00)	75.56 (6.22)	74.96 (11.28)	72.53 (9.12)
prt	99.61 (0.27)	99.74 (0.27)	98.96 (0.84)	99.68 (0.32)	99.29 (0.42)	99.55 (0.37)
lng	97.79 (2.32)	100.00 (0.00)	95.59 (3.15)	100.00 (0.00)	98.35 (2.48)	97.79 (2.32)
bcc	88.12 (3.67)	85.62 (8.12)	88.91 (5.03)	89.00 (4.76)	83.95 (4.46)	88.04 (6.54)
bco	68.29 (7.11)	96.14 (2.16)	96.19 (3.98)	96.10 (4.16)	94.19 (4.10)	96.19 (3.98)
mean	80.08	87.96	86.66	91.31	88.29	87.42

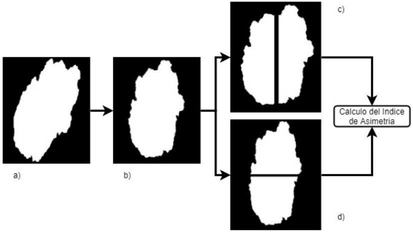
		#Features	
id	#feat.	FCBF	MSUFS
cln	2000	13.80 (1.64)	3.80 (0.84)
lym	2647	43.00 (1.41)	4.60 (0.55)
bcg	2905	36.20 (8.35)	5.00 (0.71)
prt	10935	197.20 (46.98)	2.00 (0.00)
lng	12533	107.80 (22.04)	3.60 (0.89)
bcc	22215	135.20 (9.40)	3.00 (0.00)
bco	22283	47.40 (1.46)	3.00 (1.00)
mean	–	82.94	3.57

Melanoma dataset
ABCD method



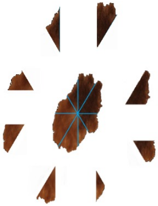
Criterion	Feature	Id	Description
asymmetry	index of asymmetry	*as*	#pixels into irregular disjoint areas
borders	segment 1-8	$b1-8$	variation of colors from center pixel to border pixels
colors	white light brown dark brown black	*wh* *lb* *db* *bk*	#pixels with tis color
dermatoscopic structures	linear branches irregular pigment network structureless areas dots and globules	*lr* *ip* *ne* *sa* *dg*	variation of distance from center to border number of unconnected pixels micro-regions number of pixel into the area number of dots and globules

Melanoma dataset
Asymmetry



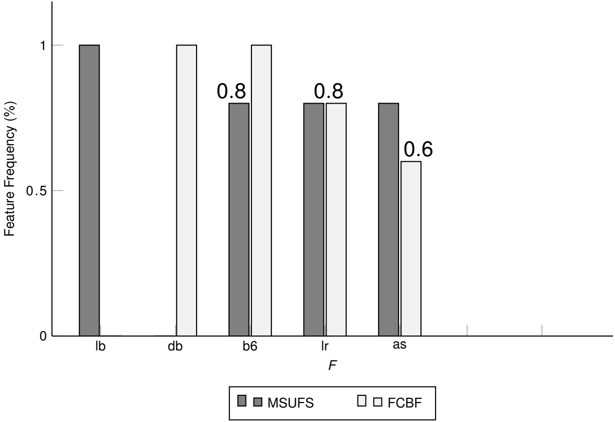
Criterion	Feature	Id	Description
asymmetry	index of asymmetry	*as*	#pixels into irregular disjoint areas
borders	segment 1-8	$b1-8$	variation of colors from center pixel to border pixels
colors	white light brown dark brown black	*wh* *lb* *db* *bk*	#pixels with tis color
dermatoscopic structures	linear branches irregular pigment network structureless areas dots and globules	*lr* *ip* *ne* *sa* *dg*	variation of distance from center to border number of unconnected pixels micro-regions number of pixel into the area number of dots and globules

Melanoma dataset
Borders



Criterion	Feature	Id	Description
asymmetry	index of asymmetry	*as*	#pixels into irregular disjoint areas
borders	segment 1-8	$b1-8$	variation of colors from center pixel to border pixels
colors	white light brown dark brown black	*wh* *lb* *db* *bk*	#pixels with tis color
dermatoscopic structures	linear branches irregular pigment network structureless areas dots and globules	*lr* *ip* *ne* *sa* *dg*	variation of distance from center to border number of unconnected pixels micro-regions number of pixel into the area number of dots and globules

id	#feat.	Accuracy FCBF	Accuracy MSUFS	#Features FCBF	#Features MSUFS
melanoma	18	79.86 ± 3.73	81.76 ± 6.11	3.4 ± 0.55	3.40 ± 0.55



- MSUFS can identify interaction among three or more features.
- MSUFS is a competitive strategy.
- The identification of multivariate interactions allow th discovery of a reduce subset of relevant features.