Evaluating the Stability of Feature Selectors that Optimize Feature Subset Cardinality

Petr SOMOL and Jana NOVOVIČOVÁ

Department of Pattern Recognition
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Prague

http://ro.utia.cas.cz

4th International Workshop on Data-Algorithm-Decision Making, 2008, Loučeň, Czech Republic

Outline

Stability of Feature Selection Algorithms

Stability of FS

for a given data set is defined

- as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution

Stability Measure

- a qualitative measure that express how much the evaluated FS process changes depending on different samplings of the same data.

Basic Notation

- $\mathcal{S} = \{\mathrm{S}_1, \ldots, \mathrm{S}_n\}$ - a system of n feature subsets
  $\mathrm{S}_j = \left\{f_k | k = 1, 2, \ldots, d_j, f_k \in \mathrm{Y}, d_j \in \{1, 2, \ldots, |\mathrm{Y}|\}\right\}$,
  $j = 1, 2, \ldots, n,\ n > 1, n \in \mathbb{N}$,
  obtained from n runs of the evaluated FS algorithm on different samplings of a given data set

- S_{id} and S_{jd} - subsets of d features, $\mathrm{S}_{id}, \mathrm{S}_{jd} \subset \mathrm{Y}$, of the same size, $0 < d < |\mathrm{Y}|$

Available Measures

- Average Normalized Hamming Distance of the system $\mathcal{S}$ (Dunne 2002)

$$ANHD(\mathcal{S}) = \frac{2}{|\mathrm{Y}|n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{|\mathrm{Y}|} |m_{ik} - m_{jk}| \qquad (1)$$

$$m_{ik} = \begin{cases} 1 & \text{if feature } f_k \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

$$0 \leq ANHD(\mathcal{S}) \leq 1$$

Available Measures

- Tanimoto Index (Coefficient)

  of similarity between two subsets S_i and S_j

  (Kalousis 2005, 2007)

$$S_K(\mathrm{S}_i, \mathrm{S}_j) = \frac{|\mathrm{S}_i \cap \mathrm{S}_j|}{|\mathrm{S}_i \cup \mathrm{S}_j|} = 1 - \frac{|\mathrm{S}_i| + |\mathrm{S}_j| - 2|\mathrm{S}_i \cap \mathrm{S}_j|}{|\mathrm{S}_i| + |\mathrm{S}_j| - |\mathrm{S}_i \cap \mathrm{S}_j|} \qquad (2)$$

$0 \leq S_K(\mathrm{S}_i, \mathrm{S}_j) \leq 1$

Available Measures

- Stability Index

  for a system $\mathcal{S} = \{\mathrm{S}_{1d}, \ldots, \mathrm{S}_{nd}\}$ for given d

  (Kuncheva 2007)

$$\mathcal{I}_S(\mathcal{S}) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_C(\mathrm{S}_{id}, \mathrm{S}_{jd}) , \qquad (3)$$

$I_C(\mathrm{S}_{id}, \mathrm{S}_{jd})$ - Consistency Index for two subsets S_{id} and S_{jd}

$$I_C(\mathrm{S}_{id}, \mathrm{S}_{jd}) = \frac{|\mathrm{S}_{id} \cap \mathrm{S}_{jd}| \cdot |\mathrm{Y}| - d^2}{d(|\mathrm{Y}| - d)} . \qquad (4)$$

$-1 \leq I_C(\mathrm{S}_{id}, \mathrm{S}_{jd}) \leq 1$

Available Measures

- Stability Measure based on Shannon Entropy

(Křížek 2007)

$$\gamma_d = -\sum_{j=1}^{K(|\mathrm{Y}|,d)} \hat{p}_{jd} \log_2 \hat{p}_{jd} \;, \tag{5}$$

$K(|\mathrm{Y}|, d) = \binom{|\mathrm{Y}|}{d}$

n_{jd} - the number of occurrences of S_{jd} in $\mathcal{S}$

$\hat{p}_{jd} = \frac{n_{jd}}{n}$ - the relative frequency of S_{jd} in $\mathcal{S}$

$0 \leq \gamma_d \leq \log(\min\{n, K(|\mathrm{Y}|, d)\})$

Novel FS Stability Measures

Novel measures for evaluating FS stability

The desirable properties of $StabMeasure(\mathcal{S})$ of the system $\mathcal{S}$:

- $0 \leq StabMeasure(\mathcal{S}) \leq 1$
- A value close to 1 implies a high level of FS algorithm stability and a value close to 0 implies a low level of FS algorithm stability

Stability measures based on feature occurrence

Basic Notation

- $\mathrm{X} \subset \mathrm{Y}$

  $\mathrm{X} = \{f | f \in \mathrm{Y}, F_f > 0\} = \bigcup_{i=1}^{n} \mathrm{S}_i, \quad \mathrm{X} \neq \emptyset$

  F_f - the number of occurrences (frequency) of feature $f \in \mathrm{Y}$ in system $\mathcal{S}$

- N - the total number of the occurrences of all features $f \in \mathcal{S}$

  $N = \sum_{g \in \mathrm{X}} F_g = \sum_{i=1}^{n} |\mathrm{S}_i|, \quad N \in \mathbb{N}, \quad N \geq n$

Consistency of the system

Consistency $C(\mathcal{S})$ of $\mathcal{S}$

the average of consistencies over all features in X:

$$C(\mathcal{S}) = \frac{1}{|\mathrm{X}|} \sum_{f \in \mathrm{X}} \frac{F_f - F_{min}}{F_{max} - F_{min}} \tag{6}$$

- $C(\mathcal{S}) = 0$ if $F_f = F_{min} = 1,\ f \in \mathrm{X}$
- $C(\mathcal{S})$ if $F_f = F_{max} = n,\ f \in \mathrm{X}$,

Weighted Consistency of System

Weighted Consistency $CW(\mathcal{S})$ of $\mathcal{S}$

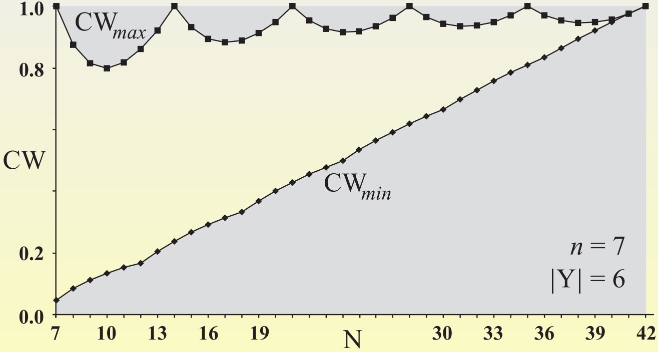
$$CW(\mathcal{S}) = \sum_{f \in \mathrm{X}} w_f \frac{F_f - F_{min}}{F_{max} - F_{min}} \tag{7}$$

$w_f = \frac{F_f}{N}$, $0 < w_f \leq 1$, $\sum_{f \in \mathrm{X}} w_f = 1$.

- $CW(\mathcal{S}) = 0$ if $N = |\mathrm{X}|$, i.e., if $F_f = 1 \;\; \forall f \in \mathrm{X}$
- $CW(\mathcal{S}) = 1$ if $N = n|\mathrm{X}|$

Weighted Consistency of System

Weighted Consistency $CW(\mathcal{S})$ Bounds



CW tends to yield the higher values the closer the sizes of subsets in system are to the size of Y – "subset-size-bias problem".

Relative Weighted Consistency of System

Relative Weighted Consistency $CW_{rel}(\mathcal{S})$ of $\mathcal{S}$

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = \frac{CW(\mathcal{S}) - CW_{min}(N, n, \mathrm{Y})}{CW_{max}(N, n) - CW_{min}(N, n, \mathrm{Y})} \qquad (8)$$

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = CW(\mathcal{S}) \quad \text{for} \quad CW_{max}(N, n) = CW_{min}(N, n, \mathrm{Y})$$

- $CW_{rel}(\mathcal{S}_{min}) = 0$
- $CW_{rel}(\mathcal{S}_{max}) = 1$

Feature Selection Experiments

Data set: from the UCI Repository

- **wine** data (13-dim., 3 classes of 59, 71, 48 samples)

Methods used:

- **Sequential Forward Selection**
- **Sequential Forward Floating Selection**
- **Dynamic Oscillating Search**

in the **Wrapper** setting that allows optimization both of

- feature subset
- subset size

FS Criterion

Classification Accuracy as FS criterion

- Gaussian classifier
- 3-Nearest Neighbor
- Support Vector Machine

- In each setup FS was repeated 1000$\times$ on randomly sampled 80% of the data (class size ratios preserved)
- In each FS run the criterion was evaluated using 10-fold cross-validation, with 2/3 of available data randomly sampled for training and the remaining 1/3 used for testing

Consistency of FS Wrappers Evaluated on Wine Data

Wrap.	FS Meth.	Classif. rate Mean	S.Dv.	Subset size Mean	S.Dv.	C	CW	CW rel	GK
Gauss.	rand	.430	.058	6.57	3.45	.505	.516	.025	.320
	SFS	.590	.023	3.73	1.70	.310	.519	.353	.379
	SFFS	.625	.023	3.58	1.23	.298	.514	.365	.389
	DOS	.636	.020	3.41	0.94	.309	.564	.453	.445
3-NN	rand	.863	.117	6.66	3.47	.511	.523	.026	.326
	SFS	.982	.004	7.12	1.47	.547	.752	.467	.615
	SFFS	.987	.003	6.91	1.60	.531	.763	.508	.637
	DOS	.989	.003	6.18	1.17	.475	.797	.643	.683
SVM	rand	.861	.125	6.40	3.50	.492	.504	.026	.307
	SFS	.980	.005	9.09	1.92	.699	.758	.203	.611
	SFFS	.989	.003	8.46	1.36	.650	.816	.516	.697
	DOS	.991	.003	7.89	1.11	.606	.841	.615	.735

Future Work

In the future we intend:

- to provide modified or simplified forms of the existing measures in a unifying framework (Dunne 2002, Kalousis 2007)
- As in the case of CW it should be possible to find bounds for the other measures and define their subset-size-unbiased counterparts, as in the case with CW_{rel}.
- to introduce an alternative approach to feature selection evaluation in form of pairwise measures that enable comparing the similarity of two feature selection processes
- the problem of very high dimensional FS stability deserves further attention as the current measures depend strongly on the d to $|\mathrm{Y}|$ ratio