ECML, 2011
Athens

Feature Selection Stability Assessment based on the Jensen-Shannon Divergence

Roberto Guzmán-Martínez
Rocío Alaiz-Rodríguez

University of León. Spain

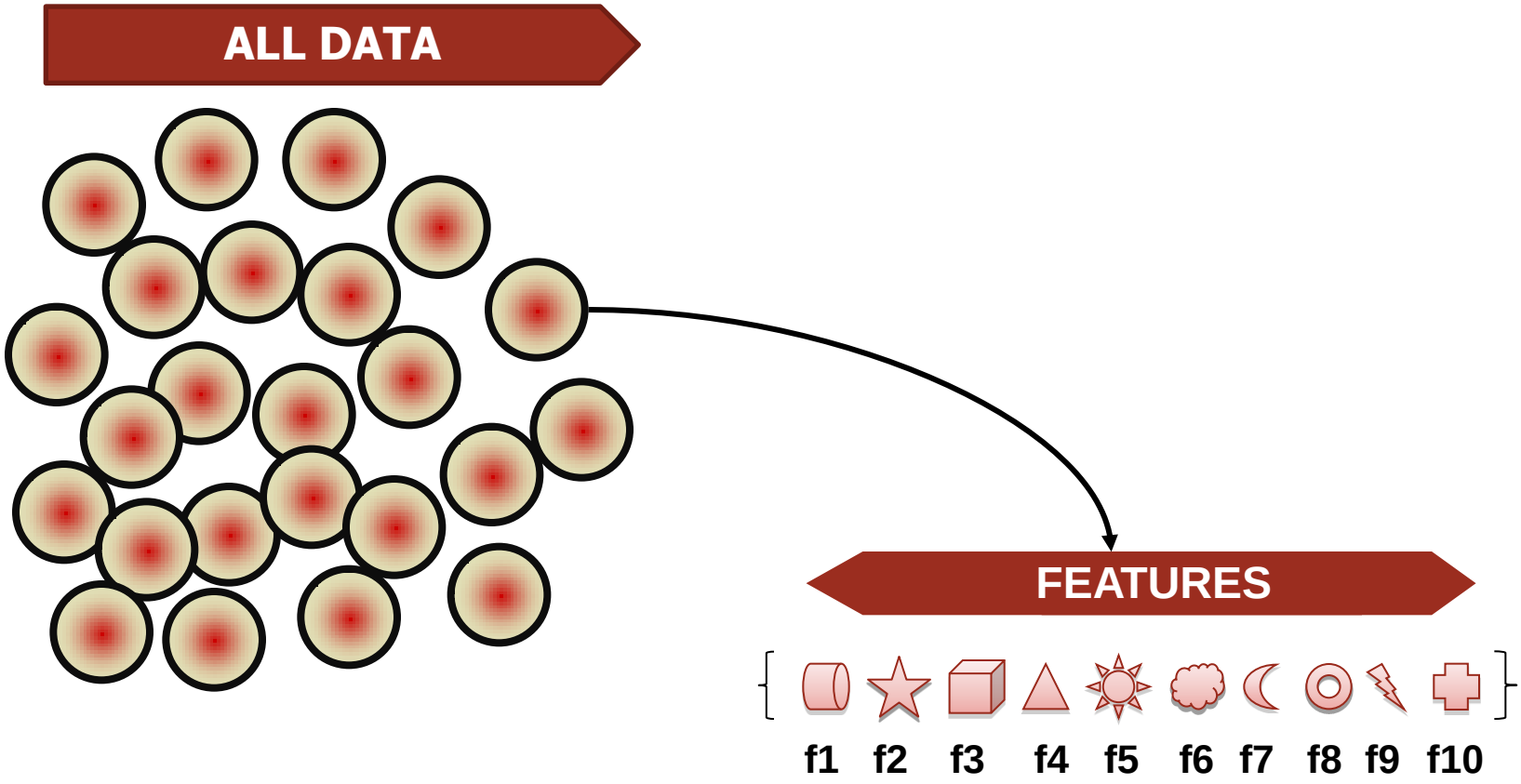
Outline

- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Outline

- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Introduction



Introduction

Feature Selection

Key stage (Multidimensional data)

Three types

- **Filter**
- **Wrapper**
- **Embedded**

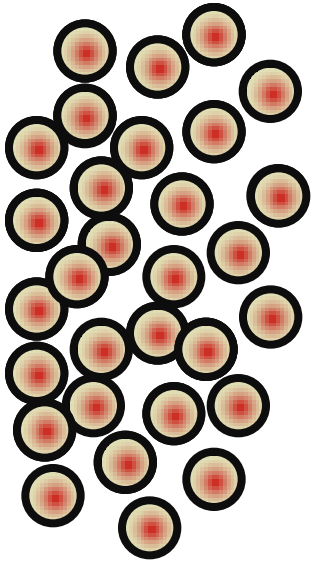
Stability: a topic of recent interest

Outline

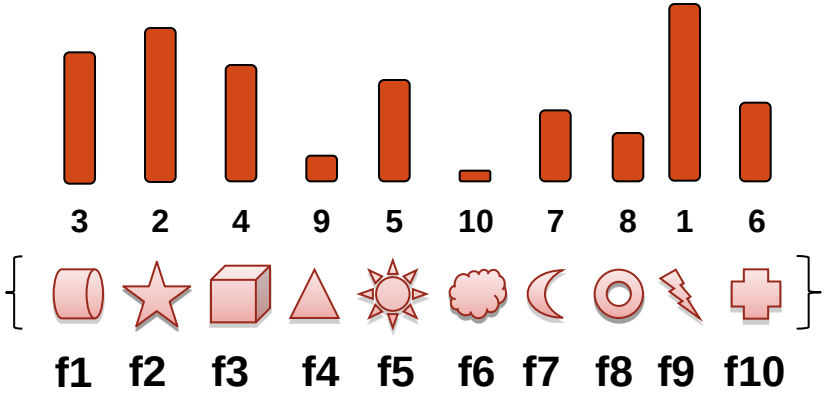
- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Feature Ranking Outcomes

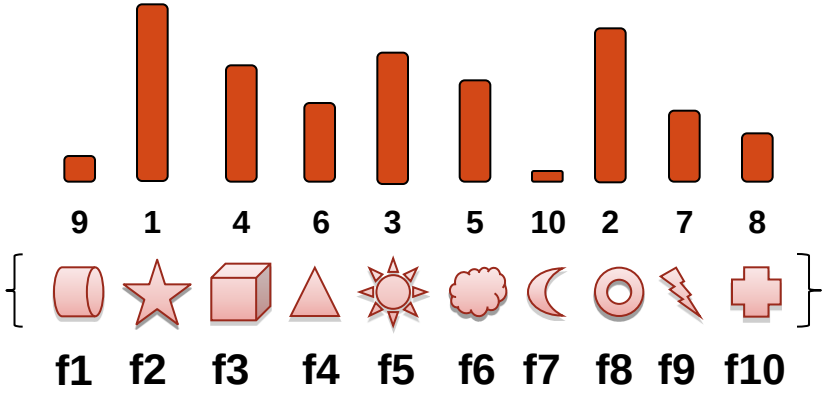
All Data



Sample 1

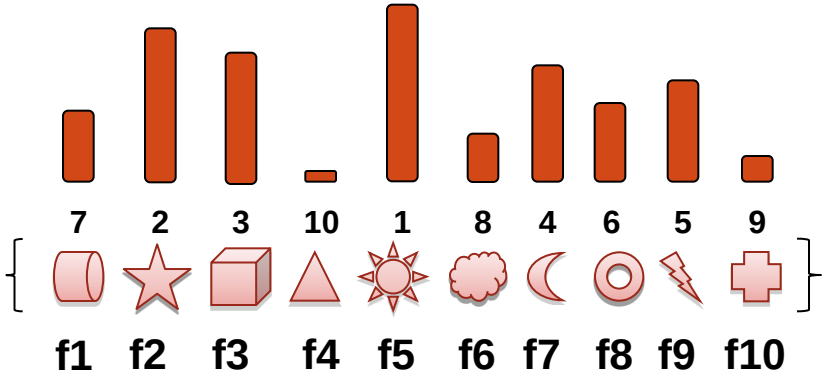


Sample 2

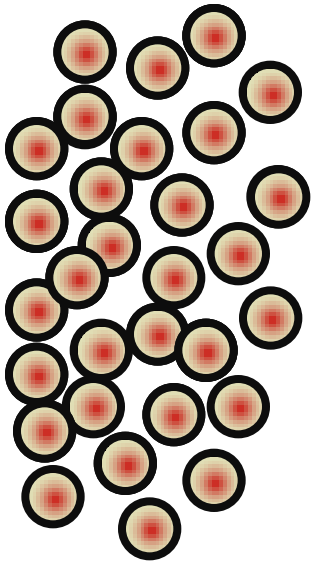


.
.
.

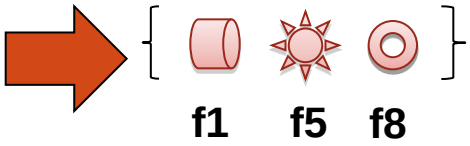
Sample k



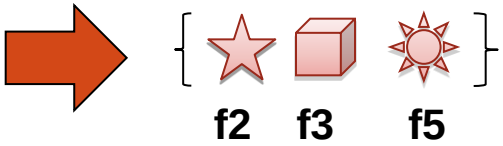
Feature Selection Outcomes



Sample 1

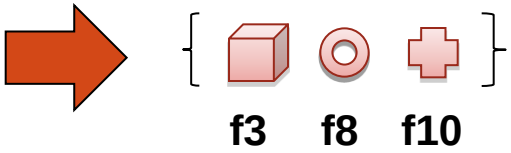


Sample 2

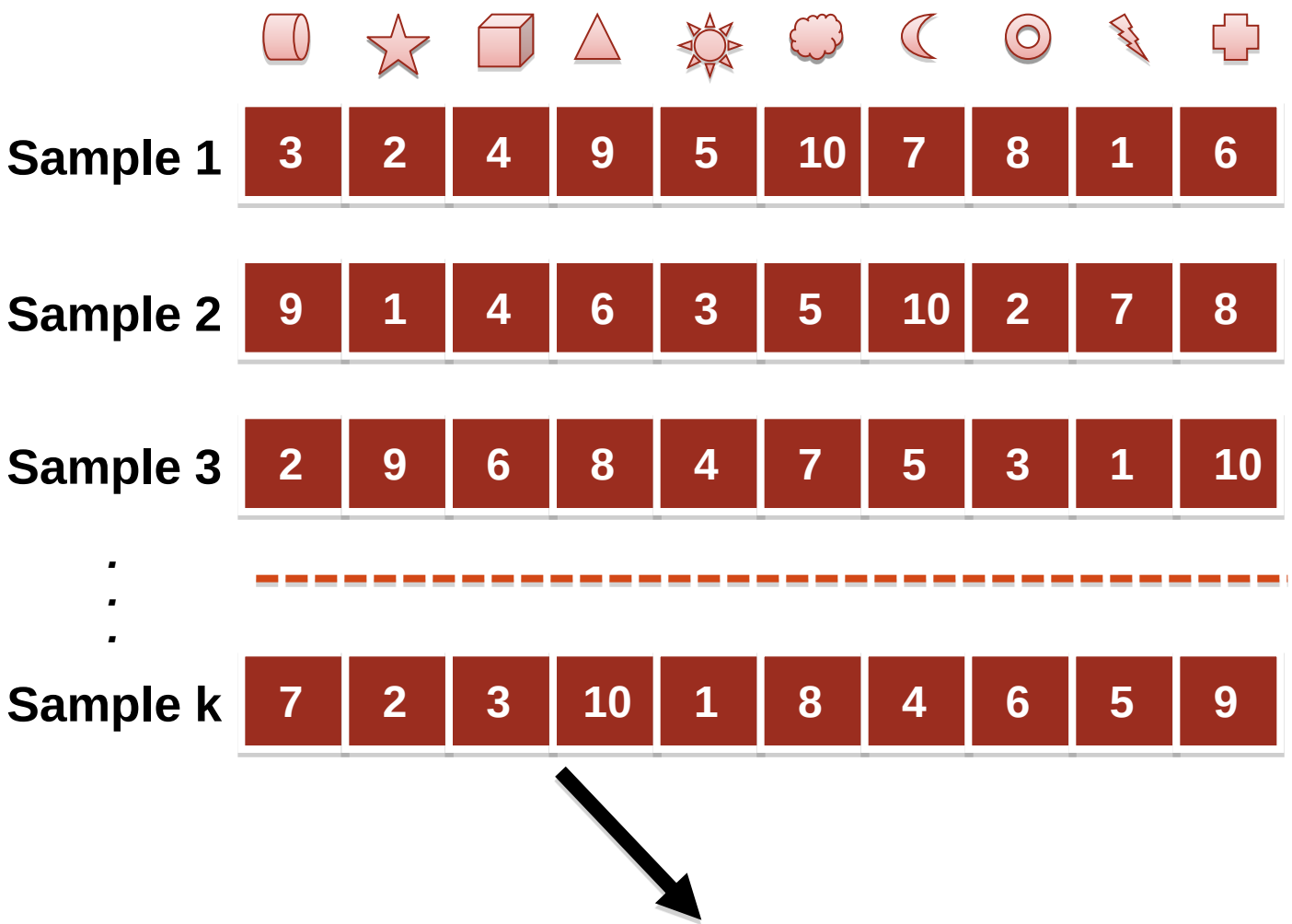


.
.
.

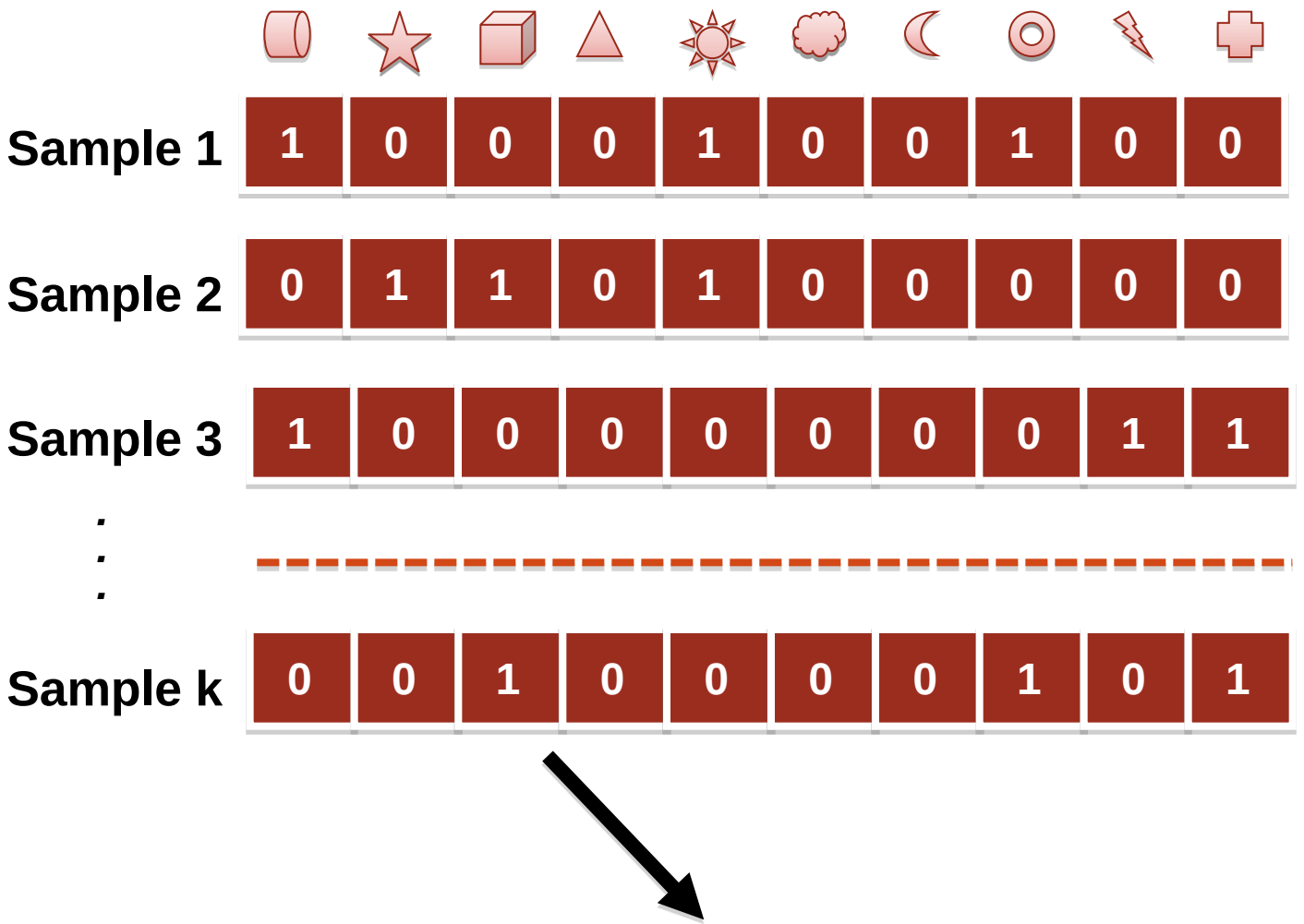
Sample k



Feature Ranking: Full Ranked lists



Feature Selection: Top-k lists



Outline

- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Feature Selection/Ranking Robustness

- Disparity among different research findings → a topic of recent interest

- Some fields require (biomedicine, bioinformatics, chemometrics):
  - Accurate classification models
  - Feature ranking (or feature subset) to better understand the data or the process

> Small variations in the data may lead to different outcomes, what makes the conclusions derived from it unreliable.

- How can we measure the stability?

Similarity between two lists (I)

Feature Ranking (full ranked lists)

Spearman's rank correlation coefficient

$$S_R(\mathbf{r}, \mathbf{r}') = 1 - 6 \sum_{i=1}^{l} \frac{(r_i - r_i')^2}{l(l^2 - 1)}$$

Manhattan distance

Feature Selection (top-k lists)

Kuncheva's stability index

$$KI(\mathbf{s}, \mathbf{s}') = \frac{ol - k^2}{k(l - k)}$$

Jaccard distance
Relative Hamming distance
Percentage of overlapping features

Similarity between two lists (II)

Partial ranked lists

- Top-k features with relative ranking among them
- Used in the domain of information retrieval to evaluate queries
- Feature importance is fundamental to carry out an analysis of the data
- No stability metrics proposed so far for these lists

Stability for a set of lists

1. Given a set of N outputs from a feature ranking/selection algorithm

$$\mathcal{A} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_N\}$$

2. Compute pairwise similarities and average the results

$$S(\mathcal{A}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_M(\mathbf{r}_i, \mathbf{r}_j)$$

Outline

- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Stability based on the Jensen-Shannon Divergence

- Metric that is able to measure the diversity among: Full ranked lists, Top-k lists and also **Partial ranked lists.**

- When the ranking is considered, differences at the top of the list would be given more importance than differences at the bottom.

- Approach based on mapping the ranking algorithm outcome **r** into a probability distribution **p**.

$$p_i = \frac{1}{2l}\left(1 + \frac{1}{r_i} + \frac{1}{r_i + 1} + \cdots + \frac{1}{l}\right)$$ [Aslam 2007]

- Similarity assessment between two ranked lists **r** and **r′**, based on measuring the divergence between **p** and **p′**.

Divergence Measures

- Kullback Leibler divergence (D_{KL})
  - Most widely used option
  - Disadvantages:
    - Asymmetric
    - It does not generalize to more than two distributions

- Jensen-Shannon divergence (D_{JS})
  - Symmetric version of D_{KL}
  - Given a set of N distributions $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$

(each one, a run of the feature ranking algorithm)

$$D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \frac{1}{N} \sum_{i=1}^{N} D_{KL}(\mathbf{p}_i || \bar{\mathbf{p}})$$

Alternatively

$$D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{l} p_{ij} log \frac{p_{ij}}{\bar{p}_i}$$

Stability based on the Jensen-Shannon Divergence

- Metric given by:

$$S_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = 1 - \frac{D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}{D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}$$

$$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \sum_{i=1}^{l} p_i log(p_i l)$$

It depends on the number of features

- Stable ranking algorithm $p_{ij} = \bar{p_i} \rightarrow D_{JS} = 0 \rightarrow S_{JS} = 1$
- Random ranking algorithm $D_{JS} = D^*_{JS} \rightarrow S_{JS} = 0$
- Any ranking outcome $S_{JS} \in (0, 1)$

Extension to Partial ranked lists

Adaptation

Mapping $$p_i = \begin{cases} \frac{1}{2k}\left(1 + \frac{1}{r_i} + \frac{1}{r_i+1} + \ldots + \frac{1}{k}\right) & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases}$$

$$S_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = 1 - \frac{D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}{D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}$$

Normalizing factor

$$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \sum_{i=1}^{l} p_i log(p_i l)$$

Extension to top-k lists

Adaptation

Mapping $$p_i = \begin{cases} \frac{1}{k} & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases}$$

$$S_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = 1 - \frac{D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}{D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}$$

Normalizing factor

$$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = log\left(\frac{l}{k}\right)$$

Outline

- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Empirical Study
Illustration on artificial outcomes (I)

FULL RANKED LISTS

- Evaluate the stability metric S_{JS} for hypothetical feature ranking algorithms

- Generation of N=100 rankings of l=2000 features

- Feature Ranking (FR). Algorithms simulated:
  - FR-0 (100 random rankings)
  - FR-1 with one fixed output,

and 99 random rankings.

  - FR-2 with two identical fixed outp

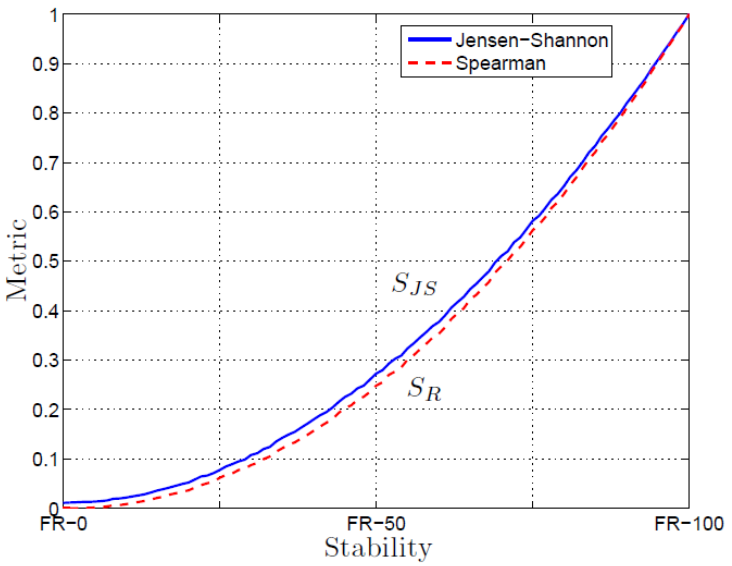
and 98 random rankings.

  - FR-i with i identical fixed outputs,

and 100 − i random rankings.

  - FR-100 with 100 identical ranking

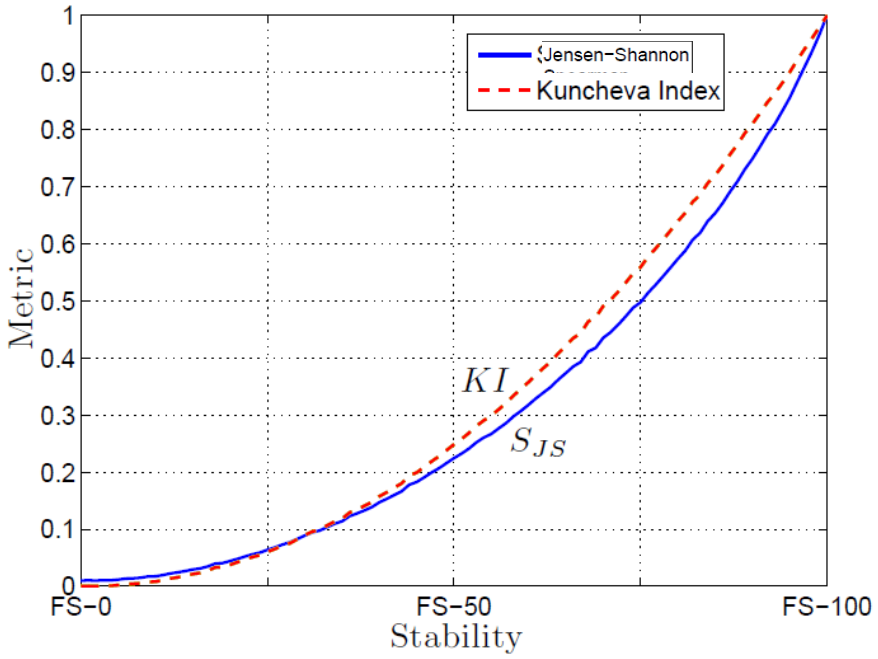
that is, an stable FR technique



Empirical Study
Illustration on artificial outcomes (II)

TOP-K LISTS

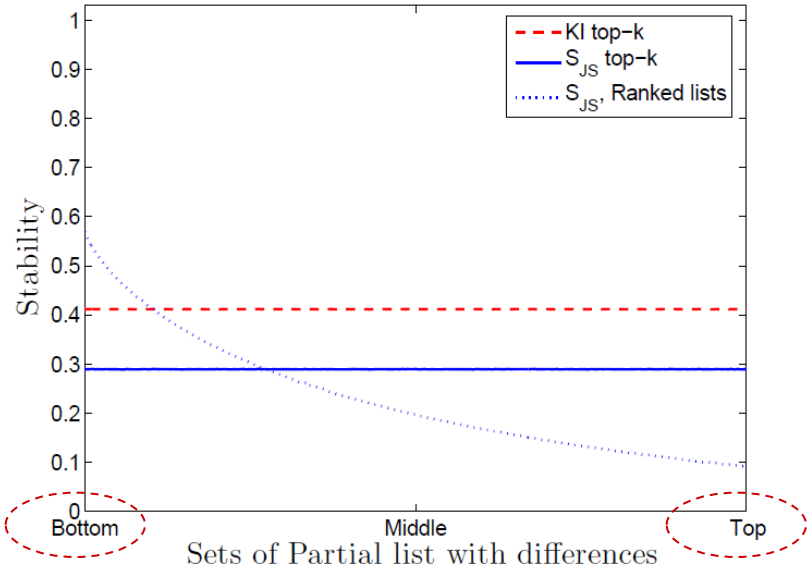
- Feature Selection (FS) algorithms simulated (obtained from FR techniques extracting the top-600 features)



Empirical Study
Illustration on artificial outcomes (III)

PARTIAL RANKED LISTS

- Sets of sublists with the 600 most important features out of 2000
- Overlap among the features is around 350

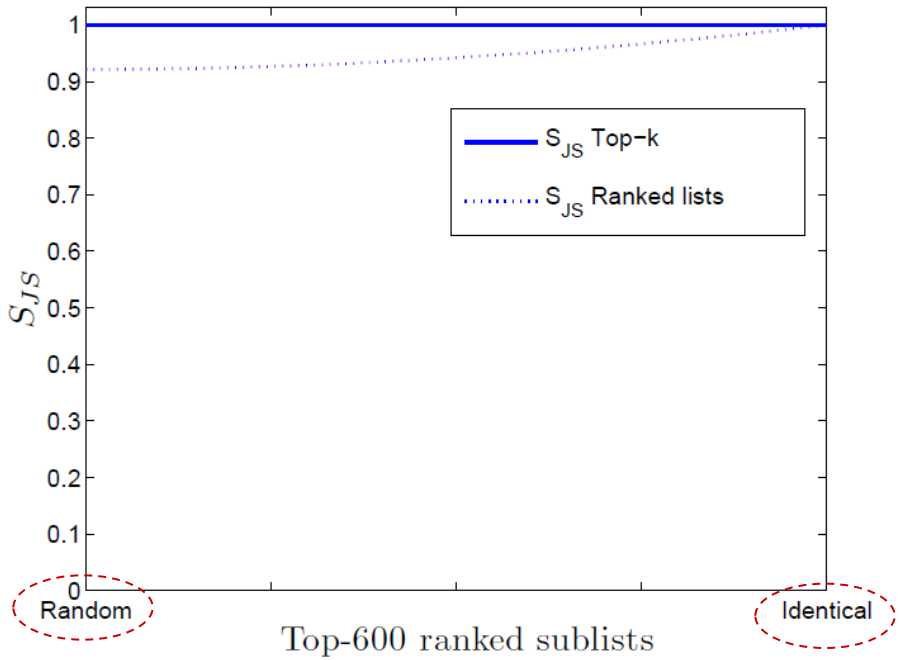


KI can not handle this information

Empirical Study
Illustration on artificial outcomes (IV)

PARTIAL RANKED LISTS

- Sets of sublists with the 600 most important features out of 2000
- Overlap among the features is 100%
- Several scenarios:



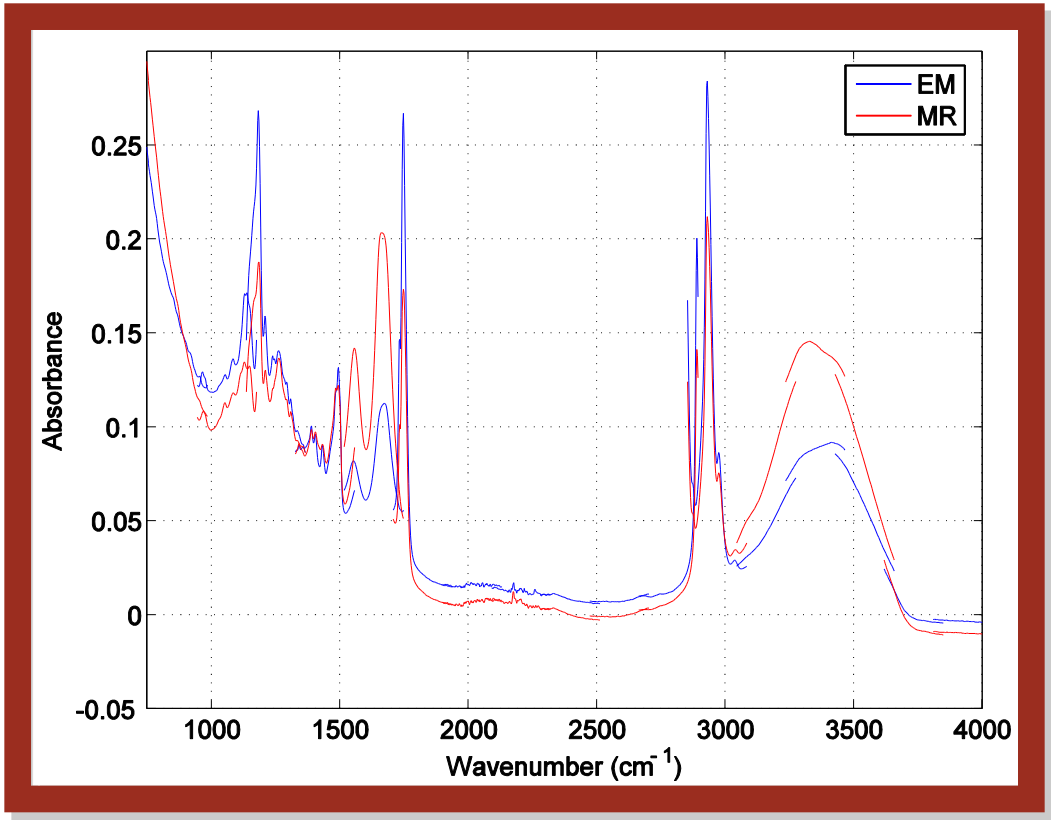
S_{JS} takes a value 0.90 when there is:

-complete agreement about the most important features

-complete disagreement about their relative importance

Empirical Study
Evaluation on Spectral Data (I)

- Assess the stability of four feature selectors.
- Experimental results were carried out with omental fat samples collected from carcasses of suckling lambs.
- Authentication of the type of feeding. (66-EM and 68-MR).
- A spectral binary dataset with 1687 features.



Empirical Study
Evaluation on Spectral Data (II)

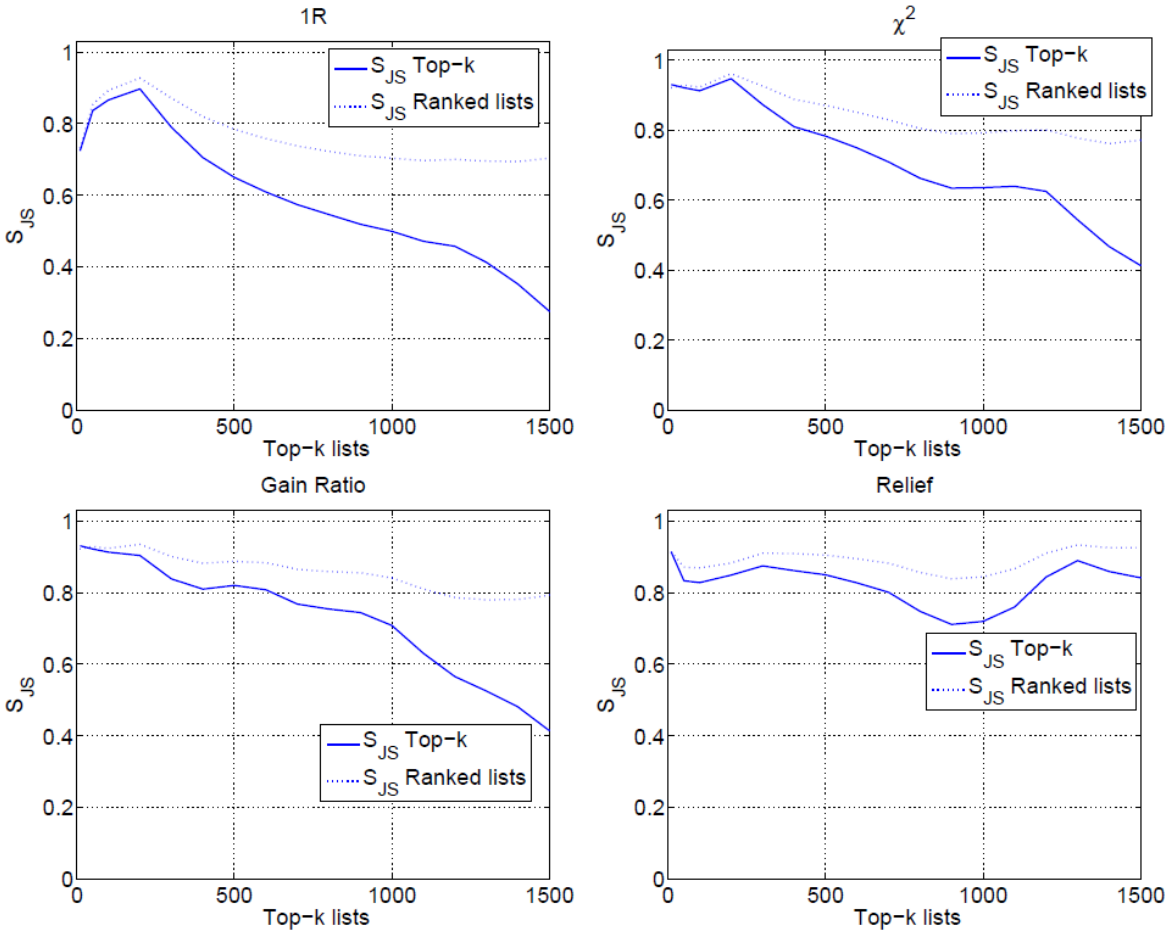
- Dataset randomly split in 10 folds.
- The feature ranking algorithm launched in 9 out of 10 (consecutive way)
- Five runs resulted in N=50 rankings
- Feature ranking (WEKA) and computation of stability with MATLAB

S_{JS} *(full ranked list)*

1R	χ^2	GR	Relief
0.87	0.92	0.94	0.94

Empirical Study
Evaluation on Spectral Data (III)

- S_{JS} (top-k) assigns a lower value. (many differences appear at the bottom of the list)
- For low values of k, the FS are quite stable.



Outline

- Introduction
- Feature Ranking/Selection
- Future Selection/Ranking Stability Metrics
- Stability based on the Jensen-Shannon divergence
- Empirical Study
- Conclusion

Conclusion

- The robustness of the feature ranking techniques used for knowledge discovery is an issue of recent interest.

- We tackle this problem and propose a metric based on the Jensen-Shannon (JS) divergence
  - Firstly,ranks are mapped to probability vectors.
  - Then, difference between vectors is evaluated using the JS divergence

  - It is able to handle: full ranked lists, top-k lists, partial ranked lists.
  - Differences at the top ranked features are penalized more
  - No need of pairwise comparisons

- The metric S_{JS} shows the relative amount of randomness of the algorithm regardless of the sublist size and evaluating directly the whole set of lists

ECML, 2011
Athens

Feature Selection Stability Assessment based on the Jensen-Shannon Divergence

Roberto Guzmán-Martínez
Rocío Alaiz-Rodríguez

University of León. Spain

Empirical Study
Evaluation on Spectral Data (IV)

- The proposed metric can be compared with the Spearman's rank correlation coefficient S_R dealing with full ranked lists
- It can also be compared with the Kuncheva's stability index KI if top-k lists are considered.
- S_{JS} is suitable for whatever problem
- Measuring the robustness with S_R or KI requires the computation of

  $\frac{50(50-1)}{2}$ pairwise similarities for each algorithm.