Università degli Studi di Milano
Master Degree in Computer Science

Information Management course

Teacher: Alberto Ceselli

Lecture 12: 21/11/2012

L. C. Molina, L. Belanche, A. Nebot
“Feature Selection Algorithms: A Survey and Experimental Evaluation”, IEEE ICDM (2002)

and

L. Belanche, F. Gonzales “Review and Evaluation of Feature Selection Algorithms in Synthetic Problems”, arXiv – available online (2011)

Feature Selection Algorithms

- Introduction
- Relevance of a feature
- Algorithms
- Description of fundamental FSAs
- Generating weighted feature orders
- Empirical and experimental evaluation

Empirical evaluation of FSAs

- First question: how do we evaluate the effectiveness of a FSA on a given dataset?
  - Relevance: features having an influence on the output
  - Irrelevance: features having no influence on the output (e.g. random values / IDs)
  - Redundance: a feature can play the role of another (e.g. strong correlation)
  - Sample size: number of tuples included in each sample by the algorithm

Scoring solutions

- Notation: $X = X_R \cup X_I \cup X_E$
  - X_R = set of Relevant features ($|X_R| = N_R$)
  - X_I = set of Irrelevant features ($|X_I| = N_I$)
  - X_E = set of rEdundant features ($|X_E| = N_E$)
  - $X^* \subseteq X$ = optimal solution
  - $A^k \subseteq X$ = solution found by the algorithm k
  - $s_X(A)$ = score: how much A and X* have in common
    - $s_X(A) = 0$ if $A = X_I$; $s_X(A) = 1$ if $A = X^*$
- Bad properties (lowering s()):
  - Relevant features lacking in A
  - Redundant features in A
  - Irrelevant features in A
- Weights α_R, α_I, α_E, can be given to these properties

Scoring solutions

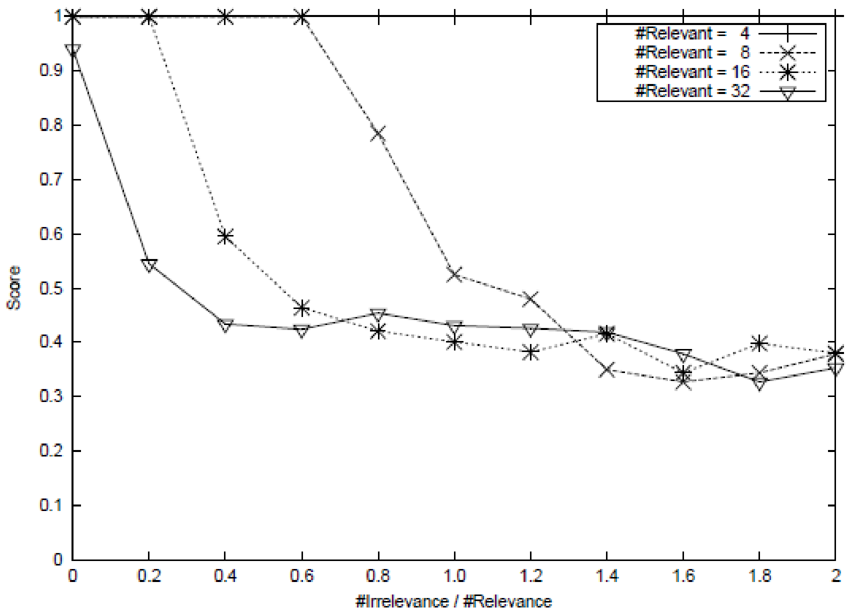
- Rough idea of the score:
  - $R = |A^k_R| / |X_R|$
  - $I = 1 - |A^k_I| / |X_I|$
  - E = ratio between the number of equivalence classes in which the original dataset is split (F) when A or X is considered (roughly speaking $E \simeq 1/|X_E| * (F(A) / F(X))$)
  - $\alpha_R + \alpha_I + \alpha_E = 1$
  - $s_X(A) = \alpha_R R + \alpha_I I + \alpha_E E$

  (for formal definition see Molina et al. 2001)
- Remark: FSAs are not optimizing the *score*!
  - FSA optimize a (local) measure of quality (e.g. consistency)
  - Results are then scored a posteriori with respect to the overall result (weighted score)

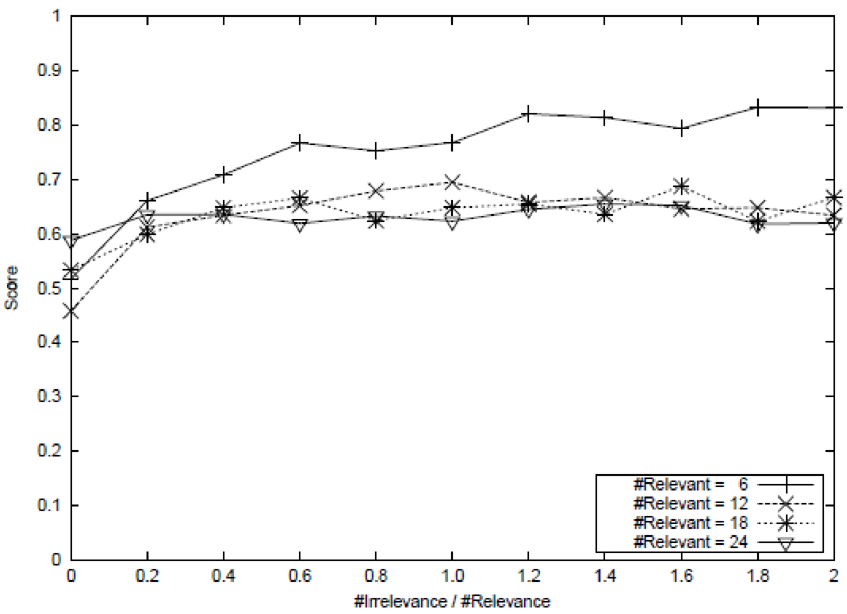
Experimental setup

- Consider three problems:
  - Parity
  - Gmonks
  - Disjunction
- Generate synthetic instances by controlling the number of relevant, irrelevant and redundant features
- Run experiments and take average values for different settings of the parameters (e.g. sample size)

Performance of FSAs



(a) Irrelevance vs. Relevance - Parity - C-SBG

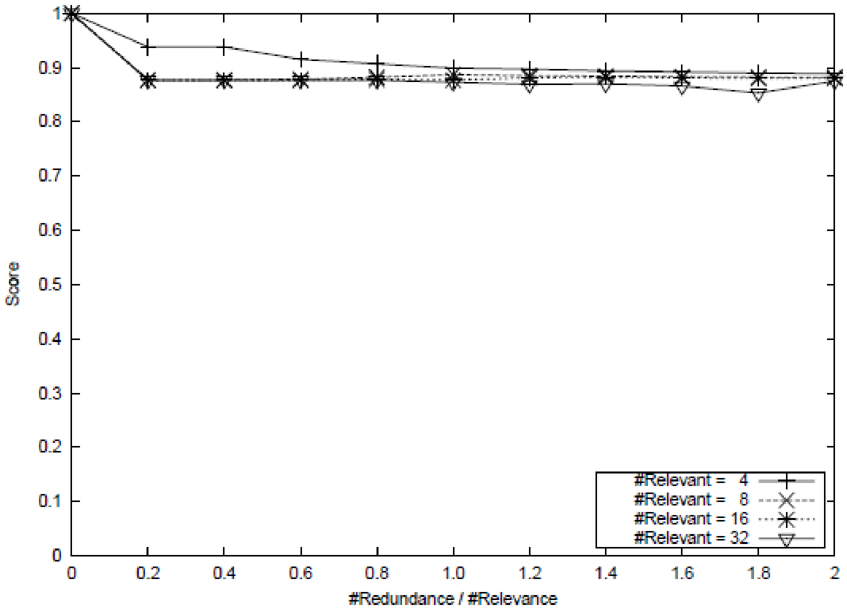


(b) Irrelevance vs. Relevance - GMonks - RELIEF

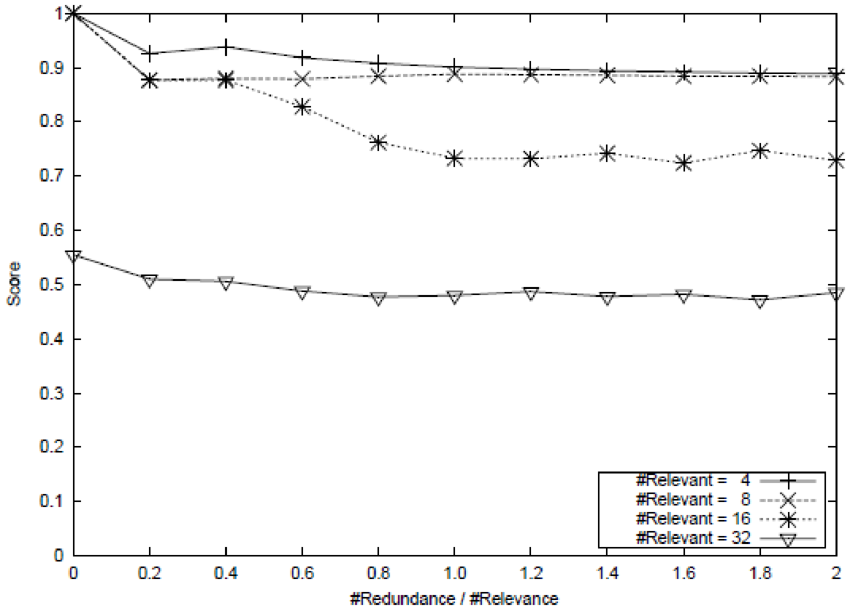
Good in the beginning, but worsens as irrelevance ratio increases

Improves as irrelevance ratio increases

Performance of FSAs



(c) Redundance vs. Relevance - Parity - LVF

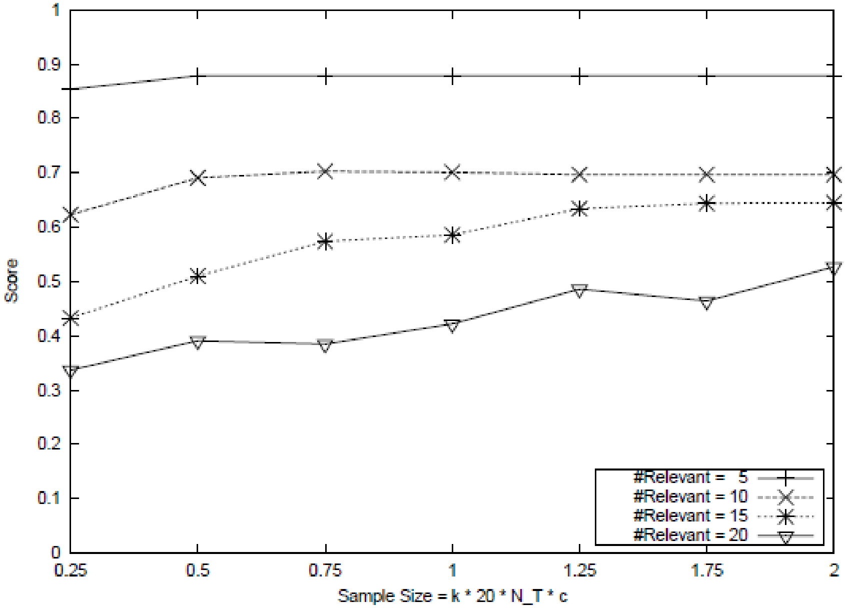


(d) Redundance vs. Relevance - Disjunction - QBB

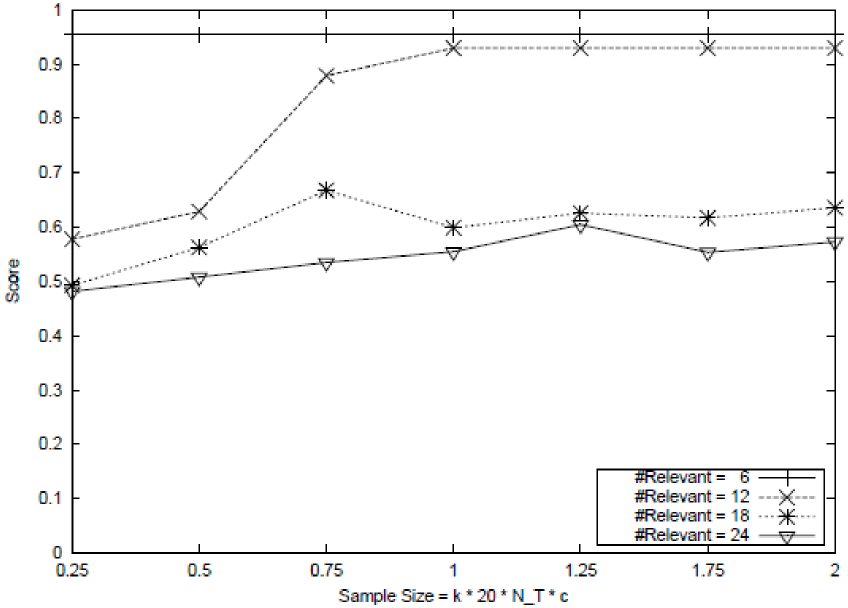
Good and stable

Very stable, but worsens as number of relevant features increases

Performance of FSAs



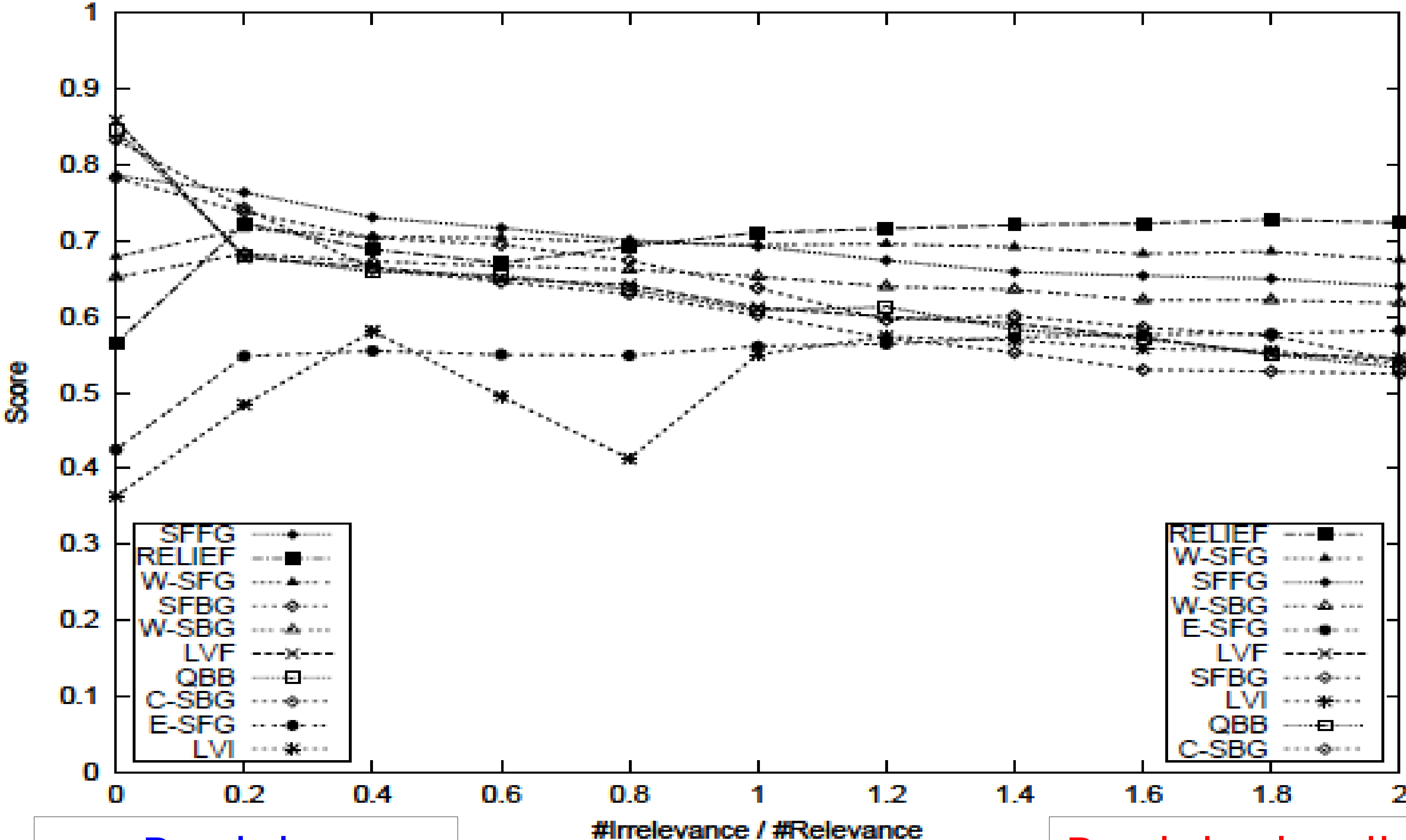
(e) Sample Size - Disjunction - LVI



(f) Sample Size - Parity - W-SBG

Curse of dimensionality effect: performance increase with sample size (more evident for higher number of relevant features)

Comparison of FSAs

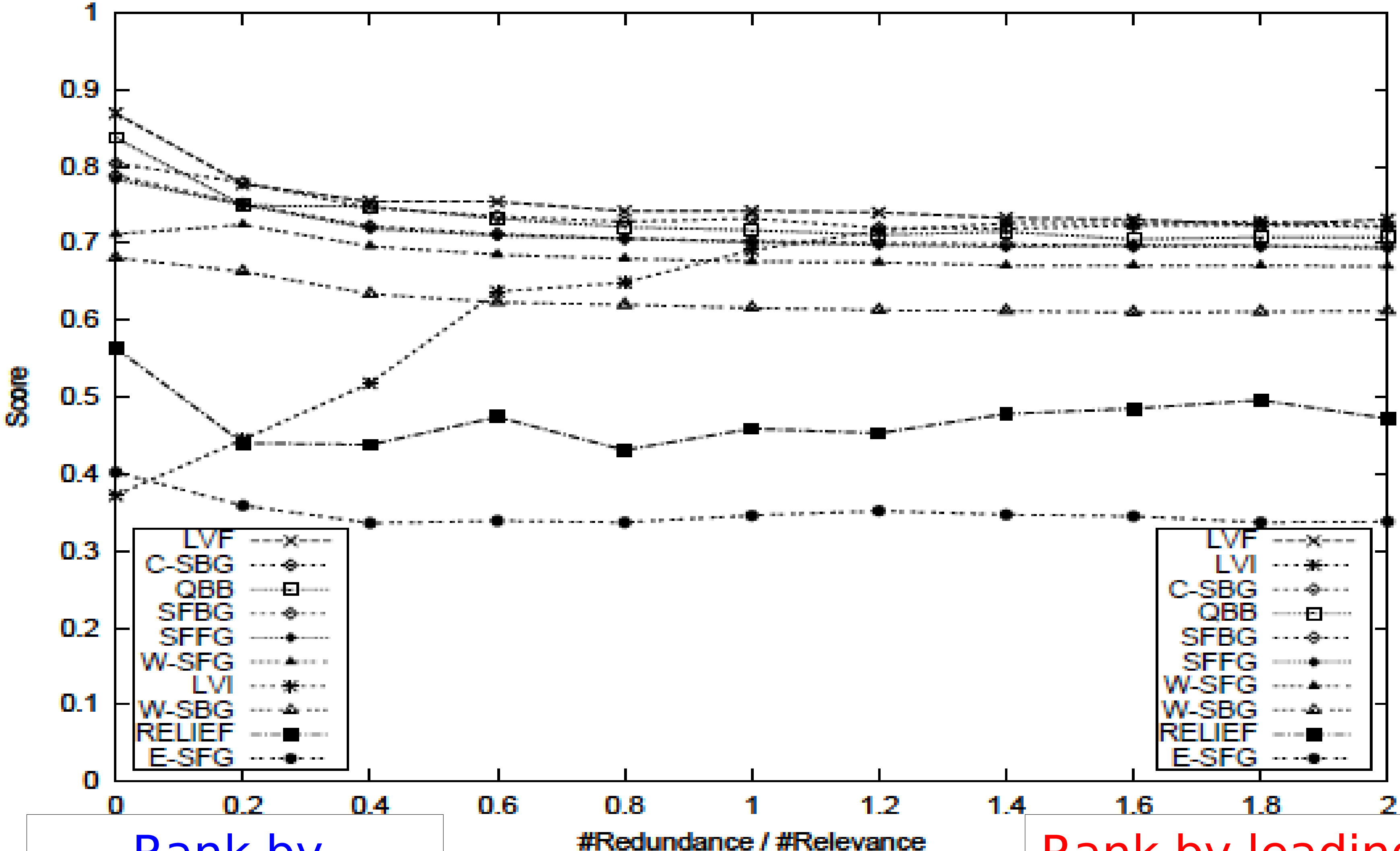


Rank by average results

(a) Irrelevance

Rank by leading in the end

Comparison of FSAs

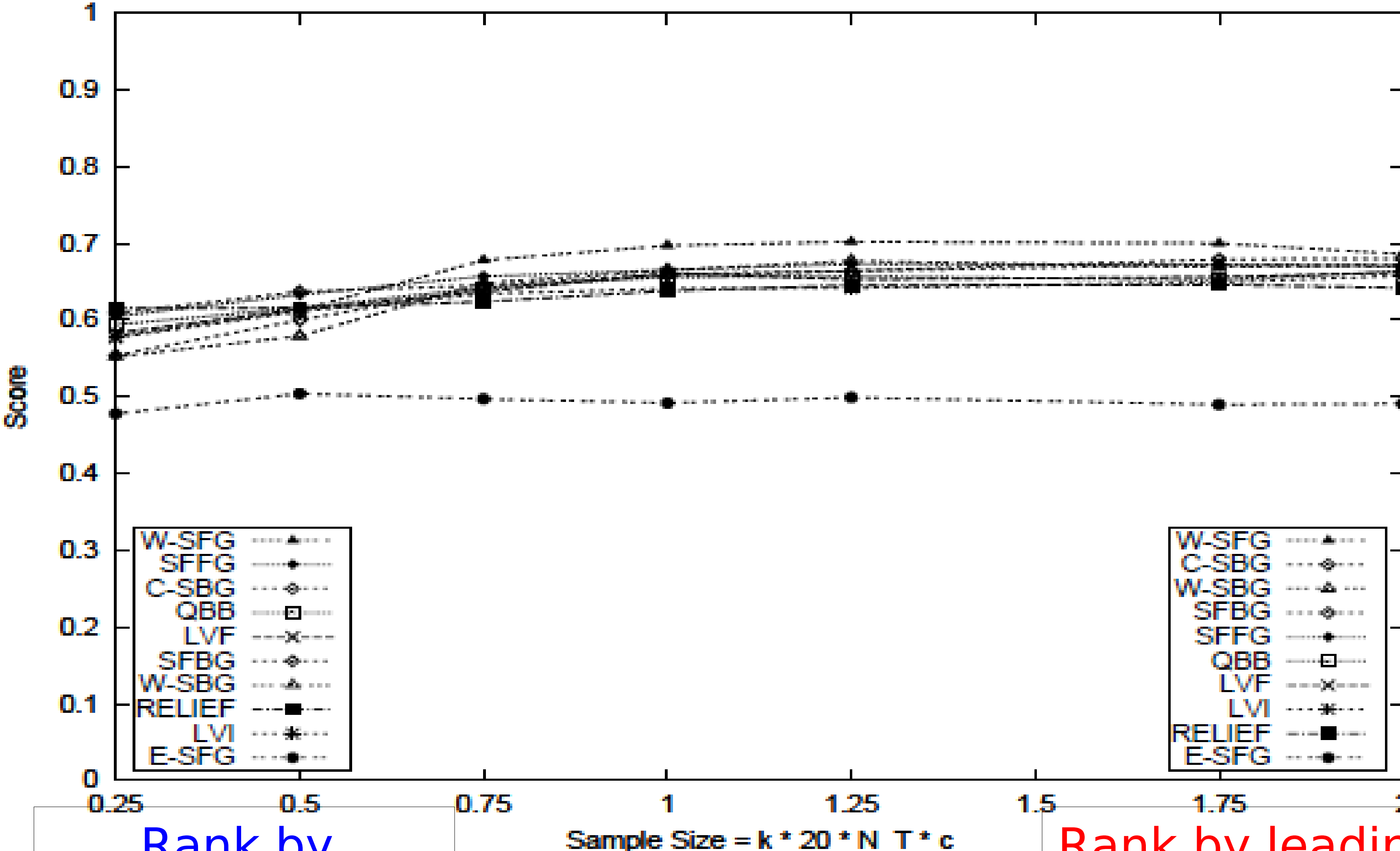


Rank by average results

(c) Redundance

Rank by leading in the end

Comparison of FSAs



(e) Sample Size

Rank by average results

Rank by leading in the end