A COMPARATIVE STUDY ON THE USE OF CORRELATION COEFFICIENTS FOR REDUNDANT FEATURE ELIMINATION

Pablo A. Jaskowiak*
Ricardo J. G. B. Campello
Thiago F. Covões
Eduardo R. Hruschka

Computer Science Department
Institute of Mathematics and Computer Science – ICMC
University of São Paulo – São Carlos - Brazil

Outline

- Introduction
- Simplified Silhouette Filter
- Correlation Coefficients
- Empirical Evaluation
- Conclusions

Introduction

- Irrelevant and redundant features
  - Feature Selection
- Possible approaches
  - Wrapper
  - Embedded
  - **Filter**

Introduction

- Simplified Silhouette Filter - SSF (Covões et al., 2009)
  - No critical user-defined parameters
  - Clustering-Based
- Correlated features
  - Clustered together
  - Eliminates redundant features

Covões, T. F., Hruschka, E. R., de Castro, L. N., dos Santos, A. M. "A cluster-based feature selection approach" in *Lecture Notes in Artificial Intelligence, v. 5572, 168-176, Springer, 2009*

Simplified Silhouette Filter

- Clustering-Based approach
  - *k*-medoids
- *k*-medoids limitation
  - *k* must be determined a priori
- Simplified Silhouette (Hruschka et al., 2006)

Hruschka, E. R., Campello, R. J. G. B., de Castro, L. N., Evolving Clusters in Gene-Expression Data, *Information Sciences, v. 176, n. 13, p. 1898-1927, 2006.*

Simplified Silhouette Filter

- Multiple runs of *k-medoids*
  - Different numbers of groups
  - Multiple runs for each number considered

- Features selected from best partition
  - SSF1 – medoid from each group
  - SSF2 – medoid from each group

    +

    feature less correlated with its medoid

Simplified Silhouette Filter

- A similarity measure must be defined
  - Correlation in this case
- Similarity choice
  - Impact in clustering algorithms
  - Impact in features selected

Correlation Coefficients

- Six different measures
  - Pearson
  - Jackknife
  - Spearman
  - Kendall
  - Goodman-Kruskal
  - Weighted Goodman-Kruskal

Empirical Evaluation

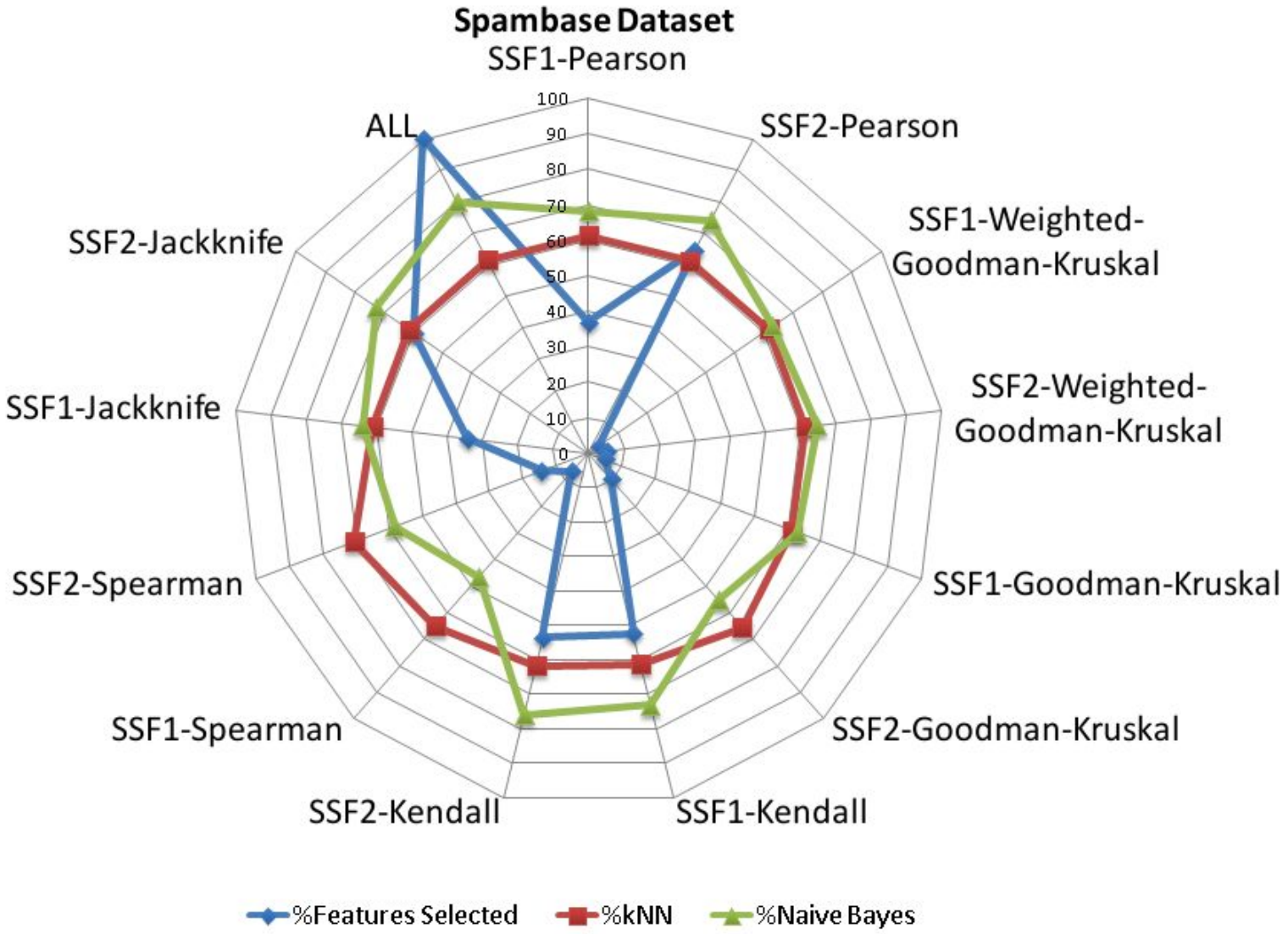
- Five datasets considered

Dataset	# Objects	# Features	# Classes
Ionosphere	351	34	2
Pima	768	8	2
Spambase	4601	57	2
Wisconsin	683	9	2
Yeast	205	20	4

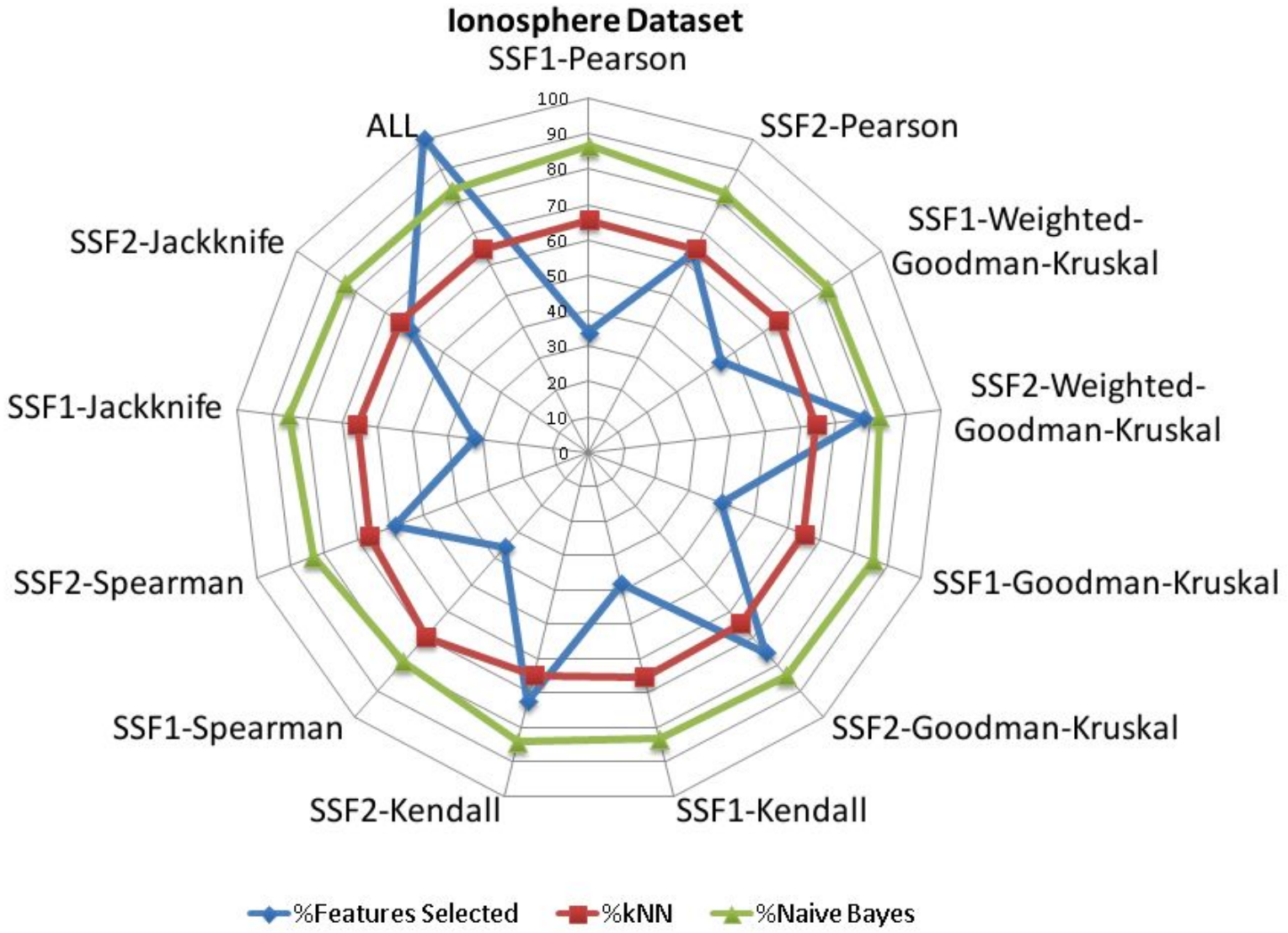
Empirical Evaluation

- Classifiers considered
  - kNN
  - Naïve Bayes
- Evaluation based on mean accuracies obtained
  - Stratified 10-fold cross-validation
  - Feature selection performed only on training set
- Number of features selected

Empirical Evaluation



Empirical Evaluation



Empirical Evaluation

- Considering SSF1
  - Same number of features selected
    - Pearson and Kendall
    - Pearson and Jackknife
  - Similar accuracies accuracies obtained
    - Spearman and Kendall
    - Pearson and Jackknife
- Considering SSF2
  - Similar accuracies obtained
    - Goodman-Kruskal and Weighted Goodman-Kruskal

Conclusions

- Considering all datasets
  - No particular correlation outperformed the others
- In some datasets interesting results were found
  - Smaller subsets
  - Better accuracies
- Correlations not commonly used in feature selection
  - Better results in some cases
- In particular studies a preliminary analysis may be interesting

Acknowledgements

Brazilian Research Agencies
CNPq and FAPESP

Questions?

Pablo Andretta Jaskowiak
pablo@icmc.usp.br

Computer Science Department
Institute of Mathematics and Computer Science – ICMC
University of São Paulo – São Carlos - Brazil