A Cluster Based Hybrid Feature Selection Approach

Pablo A. Jaskowiak and Ricardo J. G. B. Campello

Institute of Mathematics and Computer Sciences
University of São Paulo – Brazil

November 2015

BRACIS 2015 – Brazilian Conference on Intelligent Systems

Outline

- Motivation
- Simplified Silhouette Filter
- Proposed Hybrid Approach
- Results and Discussion
- Conclusions

Motivation

- Increasing data collection and storage capacities
- More objects and in most cases *more features*
  - Collect everything and decide later
- Classification task
  - Which features to use?

Motivation

- Feature Selection
  - Aims to keep *relevant* features to the problem in hand while removing *irrelevant* and *redundant* features

<u>Feature Selection</u> *vs* Feature Extraction

Feature Selection Methods

- Categorized w.r.t. their relation with the classifier

- Embedded
  - Byproduct of training
  - Model Specific
    - Decision Trees
- Filter
  - Classifier Independent
  - Ususally fast
  - Generic
- Wrapper
  - Classifier dependent
  - Usually expensive
  - Custom feature subsets
- Hybrid
  - Filter and Wrapper
  - Custom subsets
  - Moderate cost

Feature Selection Methods

- Categorized w.r.t. their relation with the classifier

- Embedded
  - Byproduct of training
  - Model Specific
    - Decision Trees

- Filter
  - Classifier Independent
  - Ususally fast
  - Generic

- Wrapper
  - Classifier dependent
  - Usually expensive
  - Custom feature subsets

- Hybrid
  - Filter and Wrapper
  - Custom subsets
  - Moderate cost

Feature Selection Methods

- Categorized w.r.t. their relation with the classifier

- Embedded
  - Byproduct of training
  - Model Specific
    - Decision Trees
- Filter
  - Classifier Independent
  - Ususally fast
  - Generic
- Wrapper
  - Classifier dependent
  - Usually expensive
  - Custom feature subsets
- Hybrid
  - Filter and Wrapper
  - Custom subsets
  - Moderate cost

Feature Selection Methods

- Categorized w.r.t. their relation with the classifier

- Embedded
  - Byproduct of training
  - Model Specific
    - Decision Trees

- Filter
  - Classifier Independent
  - Ususally fast
  - Generic

- Wrapper
  - Classifier dependent
  - Usually expensive
  - Custom feature subsets

- Hybrid
  - Filter and Wrapper
  - Custom subsets
  - Moderate cost

Feature Selection Methods

- Categorized w.r.t. their relation with the classifier

- Embedded
  - Byproduct of training
  - Model Specific
    - Decision Trees
- Filter
  - Classifier Independent
  - Ususally fast
  - Generic
- Wrapper
  - Classifier dependent
  - Usually expensive
  - Custom feature subsets
- Hybrid
  - Filter and Wrapper
  - Custom subsets
  - Moderate cost

Feature Selection Methods

- Categorized w.r.t. their relation with the classifier

- Embedded
  - Byproduct of training
  - Model Specific
    - Decision Trees

- Filter
  - Classifier Independent
  - Ususally fast
  - Generic

- Wrapper
  - Classifier dependent
  - Usually expensive
  - Custom feature subsets

- Hybrid
  - Filter and Wrapper
  - Custom subsets
  - Moderate cost

Our Approach

- Two phase feature selection method
  1. Filter
     - *Based* on Simplified Silhouette Filter
     - Redundancy
  2. Wrapper
     - Traditional wrapper approach
     - Relevance

Simplified Silhouette Filter - SSF

- Filter based on feature clustering
- Tackles feature redundancy
- Good results in comparison to competitors

Simplified Silhouette Filter - SSF

1. For *k* in 2 to k_{max}
   - Cluster features with k-medoids (repeat this *r* times)
   - Compute Simplified Silhouette (SS)
2. Select Partition with best SS
3. Select Features from partition
   - Medoid of each cluster
   - Medoid and frontier of each cluster

Simplified Silhouette Filter - SSF

- No critical parameters
  - Range for number of clusters, typically 2 to 1/2m or Sqrt(m)
  - Number of partitions for each number of clusters
  - Selection method
- No interaction with the final classifier
  - Generic feature subsets

Our Hybrid Approach

1. For *k* in 2 to k_{max}
   - Cluster features with k-medoids (repeat this *r* times)
2. Select the best partition for each *k* with SSE
3. Select Features from partition
   - Medoid of each cluster
   - Medoid and frontier of each cluster
4. Select the final subset with a wrapper
   - Feature subset with best accuracy on train set

Our Hybrid Approach

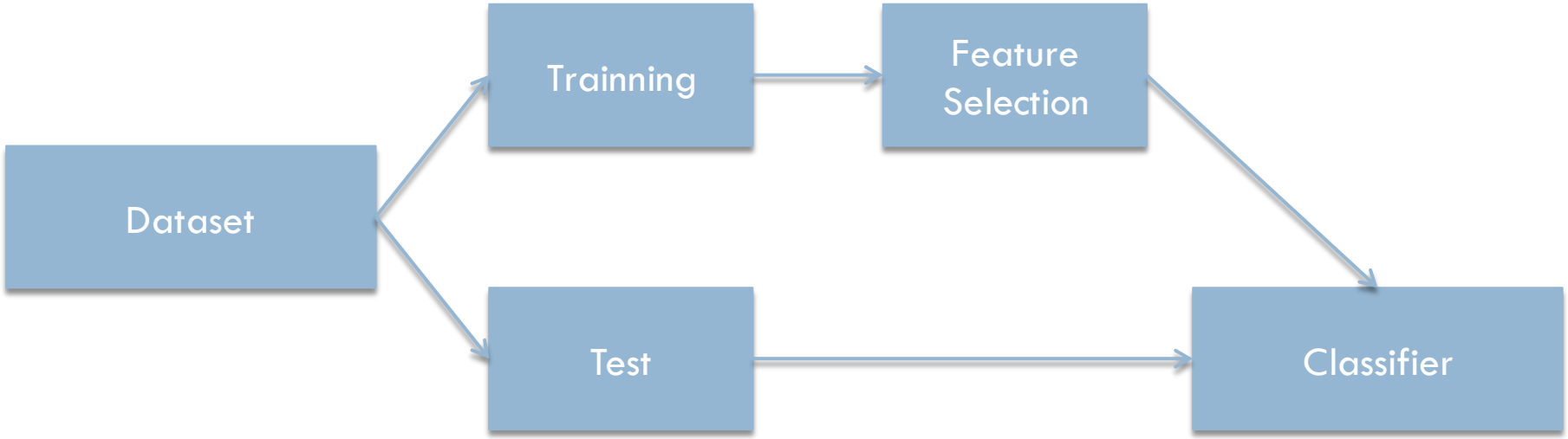
- No need of Simplified Silhouette
  - Sum of Squared Errors for fixed *k*
  - Wrapper determines the final number of features
- Wrapper examines a limited number of subsets
  - $k_{max} - k_{min} + 1$
    - 1000 Features: Sqrt(1000)-2+1 = 30 feature subsets
    - Allows the selection of a maximum of 31 features
- Same, still no critical parameters, as for SSF

Empirical Evaluation

- Two data collections
  - Collection A
    - Same datasets employed to evaluate SSF
    - 3 UCI datasets + 6 Gene expression datasets
    - From 9 to 57 features
  - Collection B
    - 35 Gene expression benchmark datasets **(de Souto et al. 2008)**
    - Around 1000 features
- Evaluated against SSF
  - Already evaluated against other methods
- Error estimates for kNN and Naive Bayes (weka default parameters)

Empirical Evaluation

- General Procedure **(Reunanem, 2003)**
  - 10 fold cross validation
- Wrapper with nested 5 fold cross validation
  - Considering only the training data!



Empirical Evaluation

- Parameters are the same for both methods
  - Pearson correlation
  - Collection A
    - Kmin = 2
    - kmax = ½ #features
    - 20 repetitions of k-medoids for each *k*
    - Both selection methods: medoid / medoid and frontier

Results on Collection A

Mean Error and Standard Deviation – Selection of One Feature per Cluster

Dataset	kNN		Naïve Bayes	
	Hybrid	SSF	Hybrid	SSF
Bio1	**00.00 ± 0.00**	02.50 ± 3.53	**00.12 ± 0.39**	02.37 ± 2.66
Bio2	**06.50 ± 2.10**	16.25 ± 5.80	**07.00 ± 2.37**	14.25 ± 3.68
Bio3	**06.50 ± 3.94**	12.75 ± 2.99	**07.37 ± 3.55**	12.37 ± 2.79
Bio4	01.00 ± 1.74	**00.25 ± 0.79**	00.87 ± 1.44	**00.37 ± 0.60**
Bio5	**01.25 ± 1.31**	02.50 ± 2.04	**00.87 ± 0.84**	02.37 ± 1.49
Spam	**11.06 ± 1.48**	14.27 ± 1.42	**21.57 ± 6.93**	24.42 ± 2.80
Wisc	**05.42 ± 2.68**	06.43 ± 2.68	**05.20 ± 2.39**	06.51 ± 2.79
Yeast	**05.40 ± 2.85**	11.16 ± 6.50	**04.91 ± 3.52**	09.45 ± 6.00
Iono	**11.67 ± 3.88**	12.53 ± 5.23	**12.82 ± 2.25**	17.38 ± 4.87

Results on Collection A

Mean Number of Features – Selection of One Feature per Cluster (Medoid)

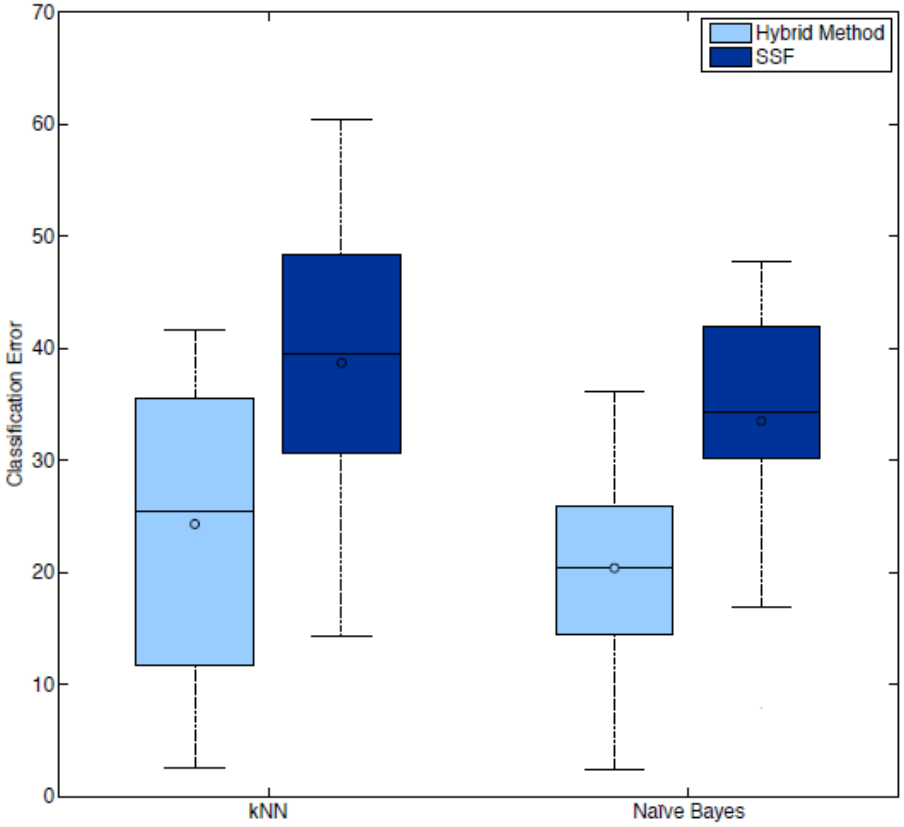
Dataset	Hybrid Approach		SSF
	kNN	Naïve Bayes	
Bio1	08.60 ± 0.69	09.30 ± 2.00	02.80 ± 0.78
Bio2	06.80 ± 1.93	07.50 ± 2.46	03.00 ± 0.00
Bio3	09.20 ± 1.39	09.90 ± 1.10	02.90 ± 0.99
Bio4	06.70 ± 1.82	07.20 ± 2.48	05.60 ± 2.36
Bio5	08.50 ± 2.36	10.90 ± 1.44	02.20 ± 0.63
Spam	26.70 ± 4.00	15.90 ± 5.76	20.80 ± 2.34
Wisc	05.80 ± 0.42	05.60 ± 0.51	02.00 ± 0.00
Yeast	10.80 ± 1.39	10.10 ± 1.37	02.00 ± 0.00
Iono	13.30 ± 4.32	16.00 ± 4.21	12.00 ± 2.10

Empirical Evaluation

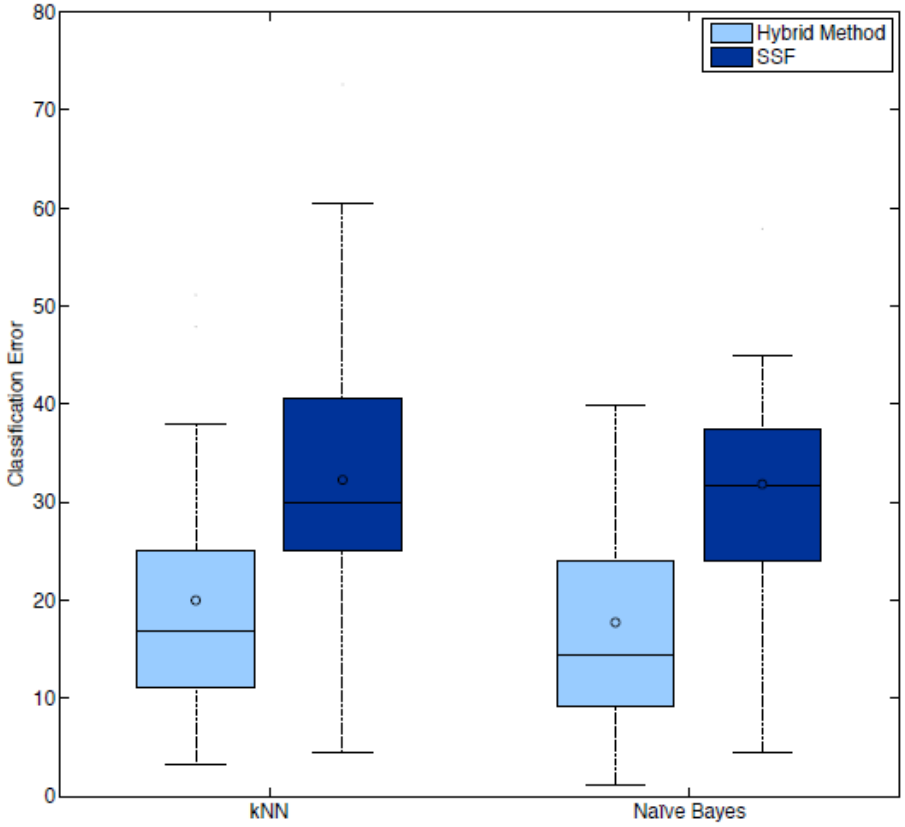
- Parameters are the same for both methods
  - Pearson correlation
  - Collection B
    - Kmin = 2
    - kmax = Sqrt(# features)
    - 20 repetitions of k-medoids for each *k*
    - Selection method: medoid and frontier

Results on Collection B

Boxplots for Error Rates



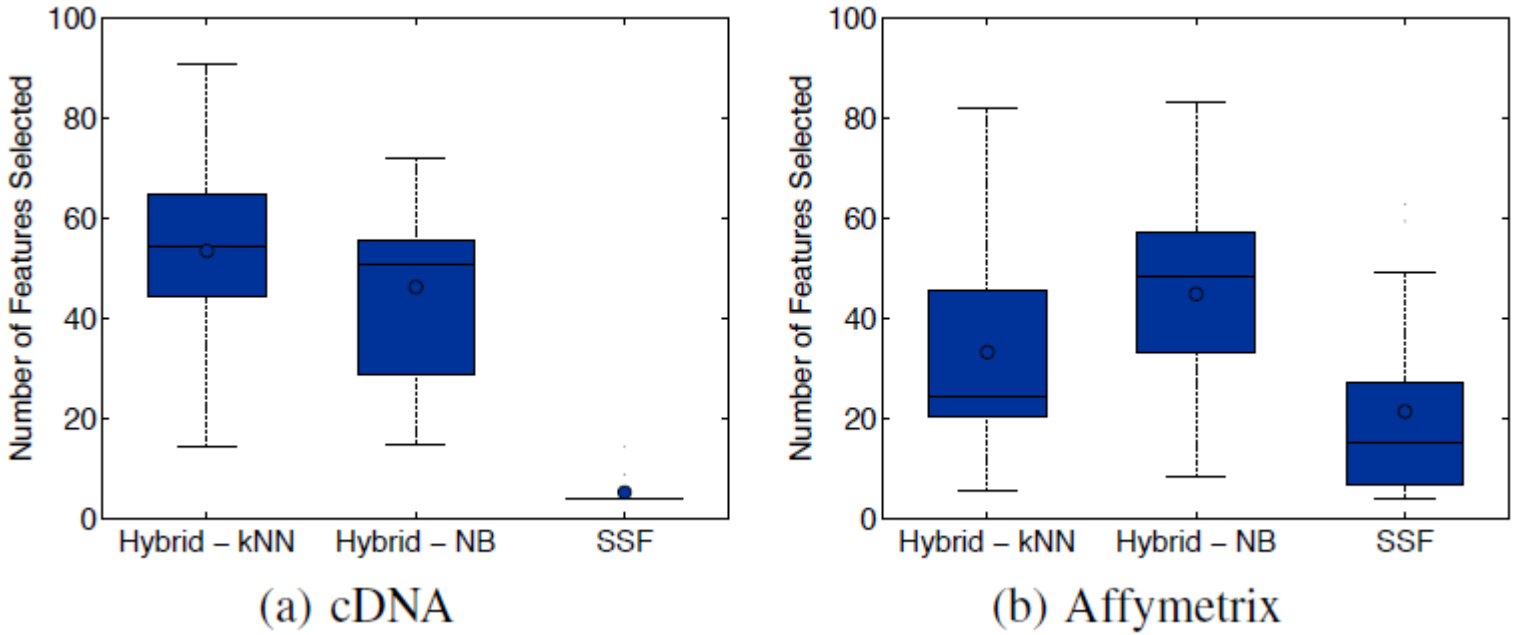
(a) cDNA



(b) Affymetrix

Results on Collection B

Boxplots for Number of Features



Conclusions

- Hybrid feature selection approach based on clustering
- Competitive results with state of the art method
- Good alternative for classification problems
  - Specific feature subsets
- Wrapper operates in a limited number of subsets
  - Considerably small number of evaluations
- Future work
  - Empirical evaluation considering running time

Acknowledgements

Brazilian Research Agency

FAPESP

Any Questions?

pablo@icmc.usp.br

Thank You!

References

T. F. Covões and E. R. Hruschka, "Towards improving cluster-based feature selection with a simplified silhouette filter," *Information Sciences*, vol. 181, no. 18, pp. 3766–3782, 2011.

M. C. P. Souto, I. G. Costa, D. S. A. Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study." *BMC Bioinformatics*, vol. 9, no. 1, p. 497, 2008.

J. Reunanen, I. Guyon, and A. Elisseeff, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, vol. 3, pp. 1371–1382, 2003.