Model Selection for Semi-Supervised Clustering

EDBT 2014

Mojgan Pourrajabi[1], Davoud Moulavi[1], Ricardo J. G. B. Campello[2] , Arthur Zimek[3] , Jörg Sander[1], Randy Goebel[1]

[1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada, [2]University of São Paulo, São Carlos, SP, Brazil,[3]Ludwig-Maximilians-Universität München, Munich, Germany

Outline

- Motivation
- Related Work
- Cross-Validation for Finding Clustering Parameters (CVCP) Framework
- Evaluation
- Conclusion

Outline

- Motivation
- Related Work
- Cross-Validation for Finding Clustering Parameters (CVCP) Framework
- Evaluation

Selecting the best model

- Challenges of using clustering algorithms:
  - Many different clustering algorithms and respective parameters
  - Parameters are difficult to set
  - Best parameter varies from one data set to another
- Can we automatically find “best” parameters using instance-level constraints?

Approaches for parameter selection

- Using relative clustering evaluation
  - Well-established criteria are only for volumetric clusters
  - Data dependent

- Semi-supervised approach
  - Focus has been only to obtain better clustering
  - Selecting appropriate parameters has not been addressed

Outline

- Introduction and Motivation
- Related Work
- Cross-Validation for Finding Clustering Parameters (CVCP) Framework
- Evaluation
- Conclusion

Evaluation of Semi-Supervised Clustering

- Internal, relative evaluation of the results
- External evaluation of the results

External Evaluation of Semi-Sup. Clustering

- ***Use all data:*** all data, including constraints, are used in evaluation [Ruiz et al. 2007, 2010, Wu et al. 2012, Zheng and Li 2011].

- ***Set aside:*** Constraints used in training stage are ignored for evaluation [Böhm and Plant 2008, Campello et al. 2013, Kestler et al. 2006, Klein and Kamvar 2002, Lelis and Sander 2009].

- ***Holdout:*** Database is divided into training and test data, then constraints are generated exclusively from training data [Law et al. 2004, Skarmeta et al. 2000].

- ***N-fold cross validation***: Data is divided into *n* folds and constraints are generated from *(n-1)* training fold combined together, this process is repeated *n* times [Basu et al. 2004, Li et al. 2009, Silva and Antunes 2012, Wagstaff and Cardie 2000, Wagstaff et al. 2001].

Outline

- Introduction and Motivation
- Related Work
- Cross-Validation for Finding Clustering Parameters (CVCP) Framework
- Evaluation
- Conclusion

Semi-Supervised Model Selection

- Goal; to provide a basis for selecting the best of parameters.

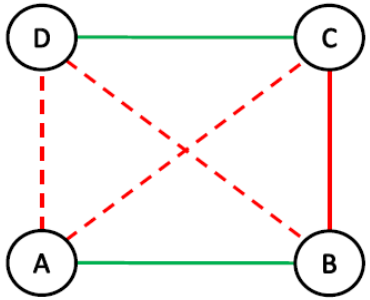
- Proposed framework:
  - Step 1: determine the quality of a parameter using n-fold cross validation by treating the partition as a **classifier** for constraints.
  - Step 2: repeat (step 1) for different parameter settings.
  - Step 3: select the parameter p* with the highest score.
  - Step 4: run the semi-supervised clustering algorithm with parameter value p* using all data as input to the algorithm.

Independence of Training and Test Folds

- Transitive and Closure of constraints causes dependence training and test set. E.g., training fold contain

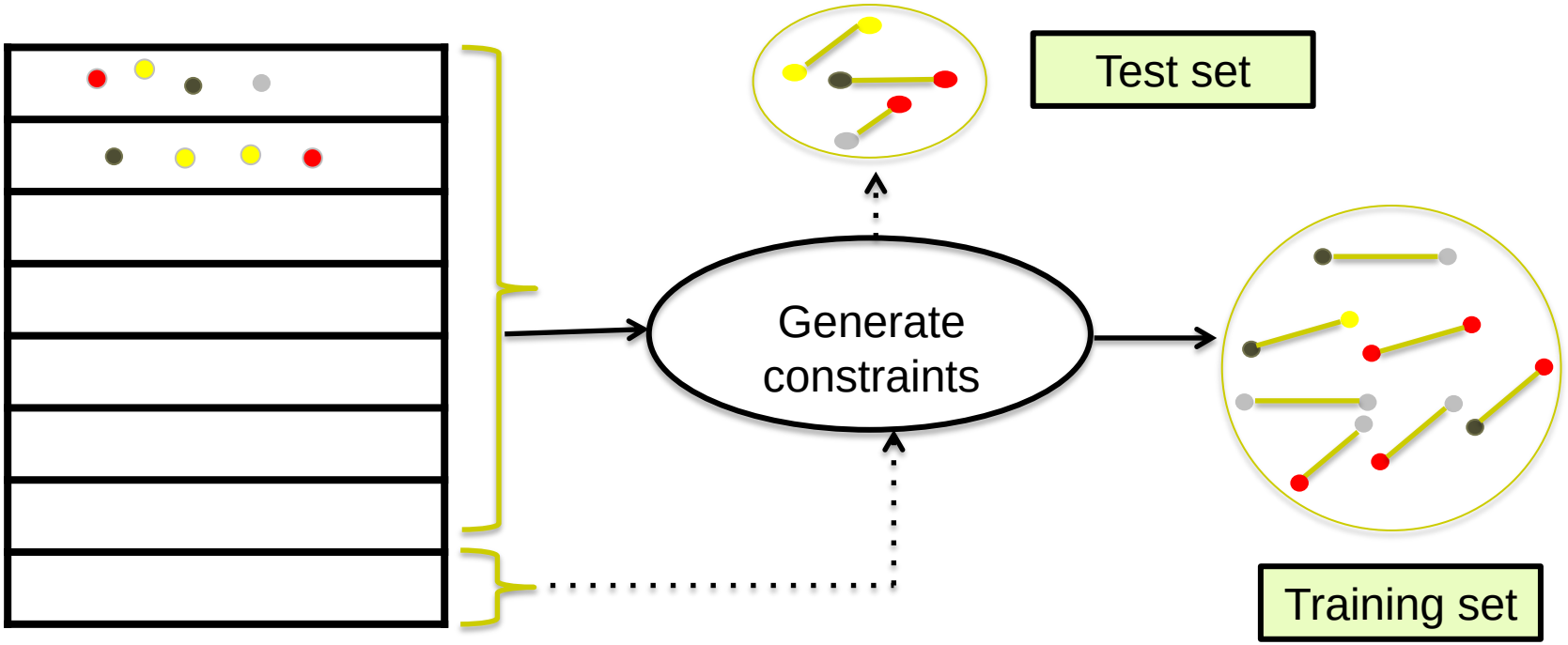
  Constraint types: Must-Link (ML), Cannot-Link(CL)

  ML(A,B), CL(B,C) and test fold contain CL(A,C)

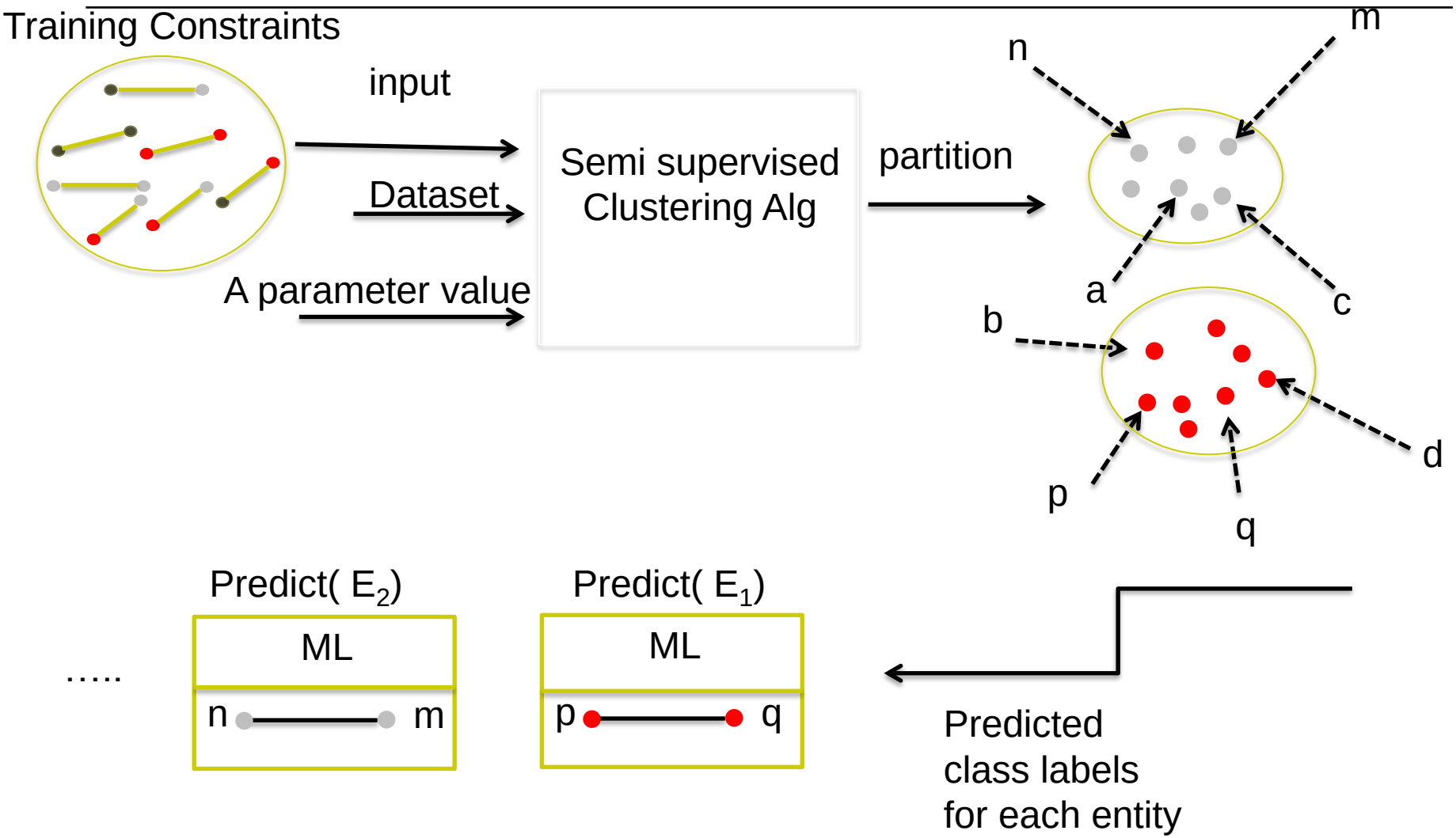


- Dealing with dependence in two possible scenarios:
  - Using label objects
  - Using pairwise instance-level constraints.

Generate the training set and test set

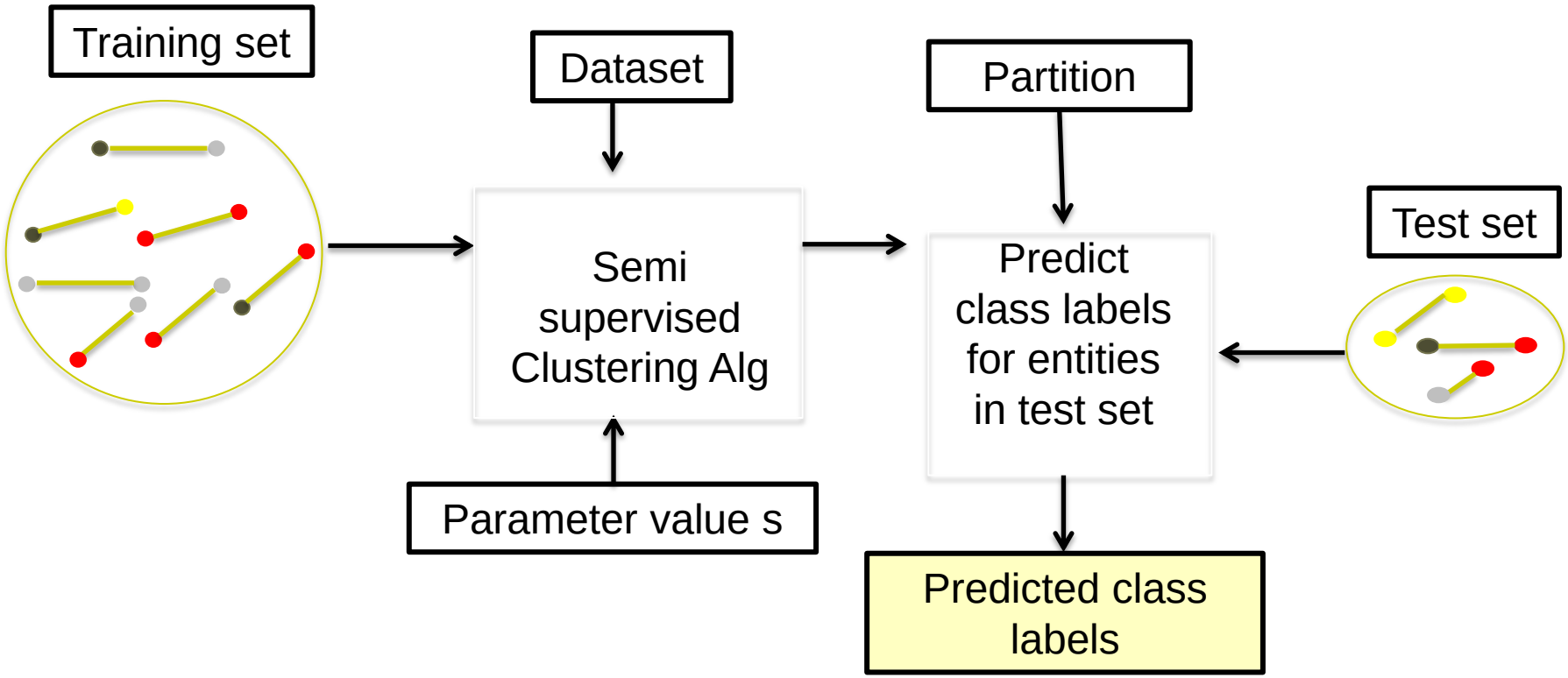


Semi-Supervised Classification of Constraints

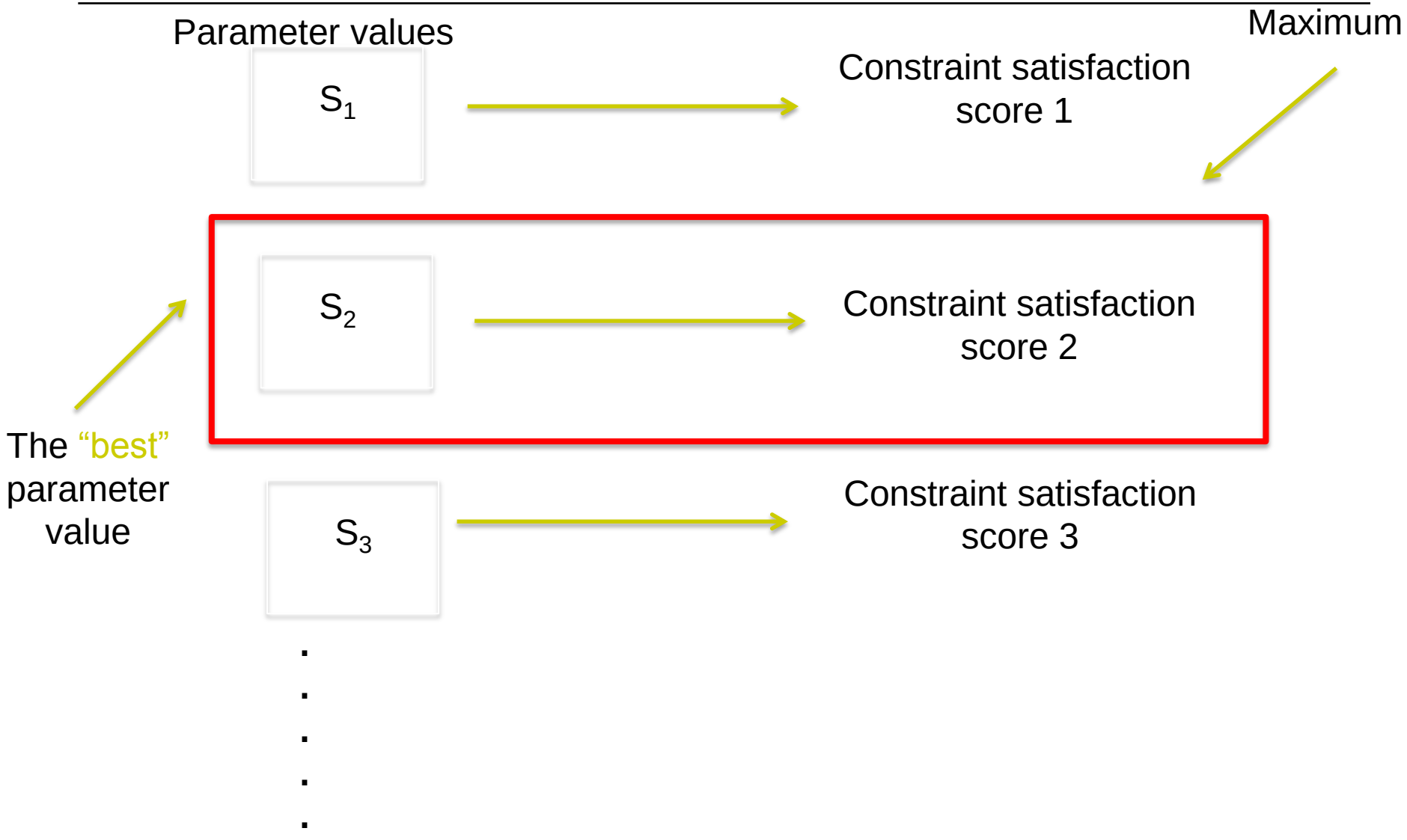


Constraint Satisfaction Score

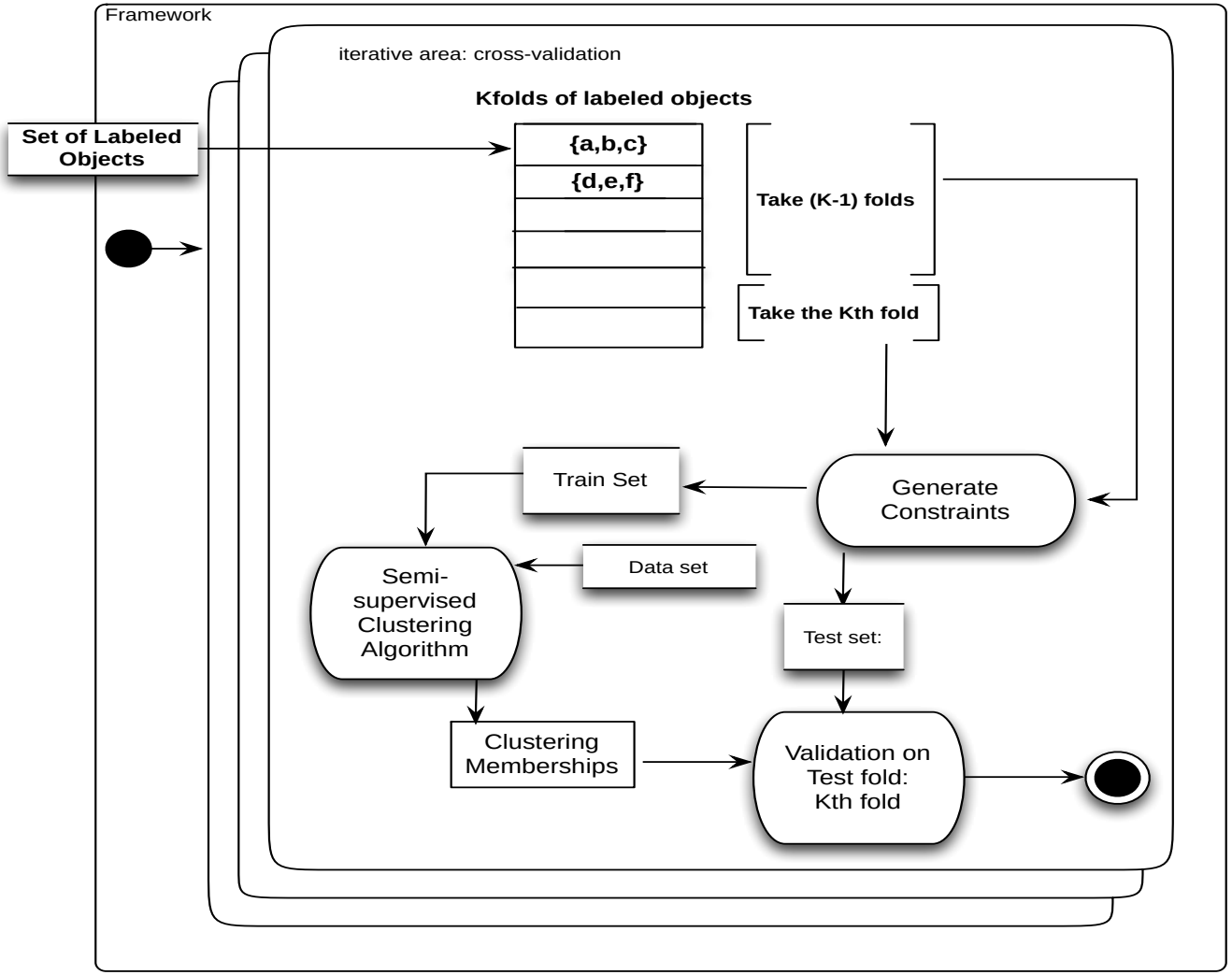
- Run the semi-supervised clustering algorithm
- Calculate the constraint satisfaction score in test fold using F-measure.



Selecting the “best” parameter



CVCP Diagram



Outline

- Introduction and Motivation
- Related Work
- Cross-Validation for Finding Clustering Parameters (CVCP) Framework
- Evaluation
- Conclusion

Evaluations

- Algorithms:
  - FOSC-OPTICSDend
  - MPCKmeans
- Evaluation
  - Evaluation I: Correlation between CVCP internal classification scores and the semi-supervised external clustering scores
  - Evaluation II: Clustering performance using the CVCP selected parameter
  - We report average results over 50 experiments for each evaluation. (In each experiment, the set of labeled points is different.)
  - We report average results for different percentage of labeled points used to produce constraints.

Evaluation I: Correlation

- Input: Data set X, a fixed set of labeled points P and a range of the parameter values S.
- For all parameter ***s:***
  - Record the constraints satisfaction score for constraints derived from P, computed within the CVCP framework for the parameter value s. (Classification score)
  - Run the semi-supervised clustering method with parameter ***s***.
  - Record clustering quality of the output partition. (Clustering score)
- Compute correlation coefficient of the classification and clustering scores.

Correlation – Label Scenario

Algorithm	FOSC-OPTICSDend			MPCKmeans		
Label / Dataset	5%	10%	20%	5%	10%	20%
ALOI	0.8019	0.9674	0.9687	0.9661	0.9237	0.9238
Iris	0.6818	0.6125	0.9902	-0.1643	0.0062	-0.3155
Wine	0.902	0.788	0.9381	0.7021	0.6639	0.2282
Ionosphere	0.9177	0.9888	0.9695	0.5735	0.4863	0.4211
Ecoli	0.688	0.8819	0.457	0.436	-0.0508	0.1017
Zyeast	0.9736	0.9433	0.9872	-0.4847	-0.7123	-0.7151

Correlation of internal scores with Overall F-Measure in label scenario

Correlation – Constraint Scenario

Algorithm	FOSC-OPTICSDend			MPCKmeans		
Constraint / Dataset	10%	20%	50%	10%	20%	50%
ALOI	0.8829	0.9013	0.9029	0.7755	0.9256	0.9314
Iris	0.7696	0.9066	0.8688	0.2755	-0.1921	-0.0486
Wine	0.797	0.8151	0.8034	0.2416	0.3136	0.2924
Ionosphere	0.9813	0.9881	0.9681	0.3021	0.5354	0.2191
Ecoli	0.945	0.9412	0.8679	0.2615	0.4875	0.391
Zyeast	0.914	0.9285	0.9081	-0.6421	-0.729	-0.6502

Correlation of internal scores with Overall F-Measure in constraint scenario

Evaluation II: Cluster Quality

- Input: Data set X, a fixed set of labeled points P and a range of the parameter values S.
- To get CVCP results:
  - Run CVCP for X using constraints derived from P to select the “best” parameter value ***s’***
  - Run Semi-supervised clustering with ***s’***
- To get expected results:
  - Run Semi-supervised Clustering for all parameter values
  - Record clustering quality of the output partition with respect to the ground truth (excluding the points in P).
  - Average all recorded clustering qualities: (expected performance).

Comparison of Clustering Quality

Algorithm	FOSC-OPTICSDend			
Dataset	CVCP Mean	Expected Mean	CVCP std	Expected std
ALOI	**0.8485**	0.7293	0.062	0.0071
Iris	**0.7615**	0.7006	0.0401	0.0066
Wine	**0.4717**	0.4569	0.0261	0.0161
Ionosphere	**0.6189**	0.5738	0.0086	0.0065
Ecoli	**0.6026**	0.5659	0.0723	0.0071
Zyeast	**0.9349**	0.8939	0.0347	0.0297

Average performance using 10 percent of labeled data as an input.
100 out 100 in ALOI dataset were significantly better.

Comparison of Clustering Quality

Algorithm	FOSC-OPTICSDend			
Dataset	CVCP Mean	Expected Mean	CVCP std	Expected std
ALOI	**0.8205**	0.723	0.0674	0.0115
Iris	**0.8541**	0.7483	0.0489	0.0261
Wine	**0.6139**	0.5469	0.0446	0.0333
Ionosphere	**0.5969**	0.5003	0.0264	0.0096
Ecoli	**0.5977**	0.5376	0.0267	0.027
Zyeast	**0.9586**	0.8923	0.0301	0.0286

Average performance using 10 percent of constraints from the constraint pool as an input. 97 out 100 in ALOI dataset were significantly better.

Comparison of Clustering Quality

Algorithm	MPCKmeans					
Dataset	CVCP Mean	Expected Mean	Silhoutte Mean	CVCP std	Expected std	Expected std
ALOI	**0.729**	0.6271	0.5881	0.041	0.0131	0.0162
Iris	0.5697	0.5676	0.4457	0.0539	0.0132	0.0122
Wine	**0.6397**	0.6367	0.3777	0.0278	0.0106	0.0124
Ionosphere	**0.6857**	0.6133	0.4605	0.0729	0.0078	0.0079
Ecoli	0.48	**0.4992**	0.3798	0.031	0.0072	0.0093
Zyeast	0.5303	0.536	0.5383	0.029	0.0087	0.0143

Average performance using 20 percent of labeled data as an input.
100 out 100 in ALOI dataset were significantly better.

Comparison of Clustering Quality

Algorithm	MPCKmeans					
Dataset	CVCP Mean	Expected Mean	Silhoutte Mean	CVCP std	Expected std	Expected std
ALOI	**0.7295**	0.6202	0.5815	0.0491	0.0052	0.006
Iris	**0.5991**	0.5644	0.4442	0.0072	0.0056	0.0049
Wine	**0.6395**	0.6452	0.3768	0.0052	0.0027	0.0034
Ionosphere	**0.7082**	0.6088	0.4594	0.0228	0.003	0.0027
Ecoli	0.5151	0.5079	0.3835	0.0993	0.0031	0.0044
Zyeast	0.5233	0.521	0.5351	0.033	0.003	0.0048

Average performance using 20 percent of constraints from the constraint pool as an input. 96 out 100 in ALOI dataset were significantly better.

Outline

- Introduction and Motivation
- Related Work
- Cross-Validation for Finding Clustering Parameters (CVCP) Framework
- Evaluation
- Conclusion

Conclusion

- We proposed a model selection method, CVCP, for semi-supervised clustering based on sound cross-validation procedure.
- The CVCP method automatically finds the most appropriate clustering parameter values (e.g., number of clusters, density-parameters) using some user-provided knowledge .
- The results show that using CVCP to select parameters can significantly improve the expected performance of a semi-supervised clustering method when appropriate parameter values often have to be "guessed".

Thank you