Exploring Multiobjective Optimization for Solving Multi-view Clustering Problem

Sriparna Saha

Humboldt Fellow
University of Mainz, Germany
AND
Assistant Professor, Indian Institute of Technology Patna, India

Joint work with Prof. Stefan Kramer and Mr. Sayantan Mitra
The corresponding work has been accepted in ACM Transactions on Knowledge Discovery from Data

February 11, 2018

Overview I

1. Multi-view Learning
   - What is Multi-view Learning?
   - Example of Multi-view Data
2. Multi-objective Optimization for Multi-view Learning
   - Motivation
   - Problem Formulation
   - Archived Multiobjective Simulated Annealing Algorithm (AMOSA)
3. Proposed Approach (Multiobjective Multiview Based Approach)
4. Experimental Results
   - Results on UCI Datasets
   - Results for Search Result Clustering
5. References

What is Multi-view Learning? [1]

- Learning from multi-view data by considering the diversity of different views
- Views are obtained from multiple sources or different feature subsets
- Example 1: a person can be identified by face, finger- print, signature or iris with information obtained from multiple sources
- Example 2: An image can be represented by its color or texture features, which can be seen as different feature subsets of the image.
- A function is used to model a particular view of the input data.
- Functions are jointly optimized to exploit different views of the same input data.

Example of Multi-view Data

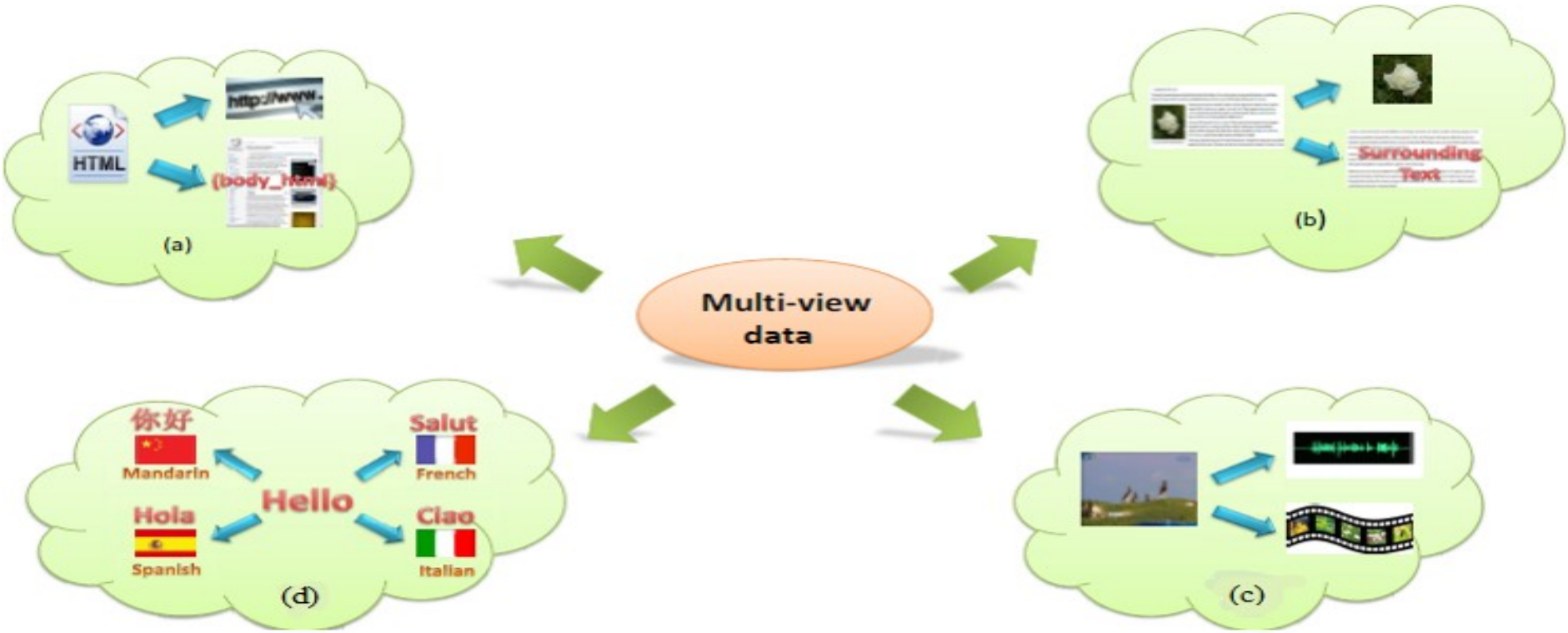


Figure: Multi-view data [1]: a) a web document can be represented by its url and words on the page, b) a web image can be depicted by its surrounding text separated by the visual information, c) video clips are combinations of audio signals and visual frames, d) multilingual documents have different views for different languages.

Disadvantages of Single-view Learning

- Conventional machine learning algorithms: concatenate multiple views into one single view to adapt to the learning setting.
- Each view has a specific statistical property.
- Concatenation causes over-fitting in case of a small sized training data.

Motivation

- Existing mult-view clustering approaches mostly categorized into 2 groups: centralized and distributed.
- Multi-view clustering approaches: a single objective function is utilized to capture goodness of a partitioning with respect to different views;
- Application of Multiobjective Optimization is limited in multi-view clustering;
  - Existing approaches assume no of clusters;
  - Some can not handle more than 2 views;
- No existing automatic multi-view multiobjective clustering technique
- Development of a partitional approach satisfying multiple objective functions + multiple views is required

Problem Formulation I

- The multi-view clustering problem is formulated as a multiobjective optimization problem:
  - Let a particular data set contain n samples each having d number of different features. $S = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n\}$
  - Total m different views
  - Get a consensus partitioning (U) satisfying all views
    - The set of data-points, S is divided into K clusters, $\{U_1, U_2, \ldots, U_K\}$
    - $U_i = \{\overline{x}_1^i, \overline{x}_2^i, \ldots, \overline{x}_{n_i}^i\}$
    - $$\cup_{i=1}^{K} U_i = N.$$

      and

      $$\cap_{i=1}^{K} U_i = \emptyset.$$

Problem Formulation II

- which simultaneously optimize some objective functions:

$$CV_1, CV_2, \ldots, CV_m, AI$$

  Here CV_i is a cluster validity index measured on the partitioning obtained after considering only view i for the given data set;
  AI measures the agreement between the partitionings obtained for different views;
- Here we have considered $CV_i = PBM$ [2] or $CV_i = XB$ [3].
- Simultaneous optimization of these objectives provides a Pareto optimal front;

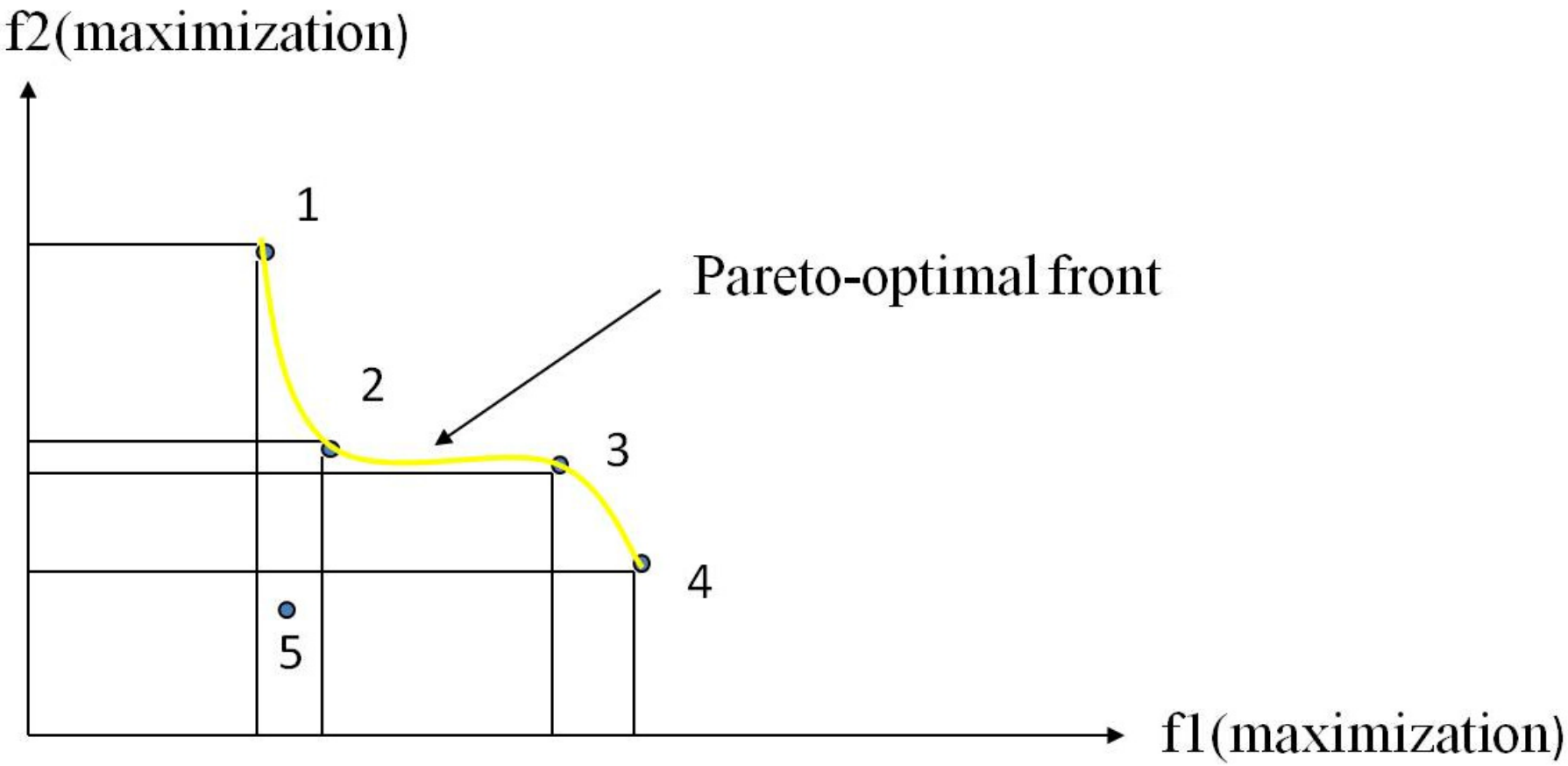


Figure: Dominance and Pareto-optimality

- Here solution 1, 2, 3 and 4 are non-dominating to each other.
- 5 is dominated by 2, 3 and 4, not by 1.

Archived Multiobjective Simulated Annealing Algorithm (AMOSA)

- AMOSA [4]
  - Based on Simulated Annealing.
  - Incorporates the concept of an archive where the non-dominated solutions seen so far are stored.
  - Uses clustering to restrict the size of the archive and to ensure diversity.
  - Uses amount of domination for computing the acceptance probability depending on domination status of the new solution, current solution and archive
- Two limits kept on the size of the archive: Hard-limit and Soft-limit.
- During the process the non-dominated solutions are stored in the archive as and when they are generated until the size of the archive increases to Soft-limit.
- If more non-dominated solutions are generated, the size of the archive is first reduced to Hard-limit by applying clustering.

AMOSA Algorithm

```
Input: Tmax, Tmin, Hard-limit, Soft-limit, iter, alpha.
Set temp=Tmax.
Initialize the archive.
Current-pt=random(archive).
While (temp> Tmin){
   for(i=0; i<iter; i++){
     New-pt=Perturb(Current-pt)
     Decision about Current-pt based on domination status
     (many subcases depending on the position of current-pt,
     new-pt and points of the archive)
   }
   temp=alpha*temp
}
If Archive-size > Soft-limit
      Archive=cluster (Archive, Hard-limit)
Output Archive
```

String Representation I

- Let us assume that each sample point has m different views, V_1, $V_2, \ldots, V_m$.
- The distribution of features in m different views is available, $\sum_{i=1}^{m} |V_i| = d$
- Archive member i represents the centroids of K_i clusters and has length l_i, where $l_i = d * K_i$.
- $$K_i = (rand() \bmod (K_{max} - 1)) + K_{min} \quad (1)$$

$C_1^1(1)$	$C_1^1(2)$	$C_1^1(3)$	$C_1^2(1)$	$C_1^2(2)$	$C_2^1(1)$	$C_2^1(2)$	$C_2^1(3)$	$C_2^2(1)$	$C_2^2(2)$

Figure: Example of a string representation.

Formation of Clusters

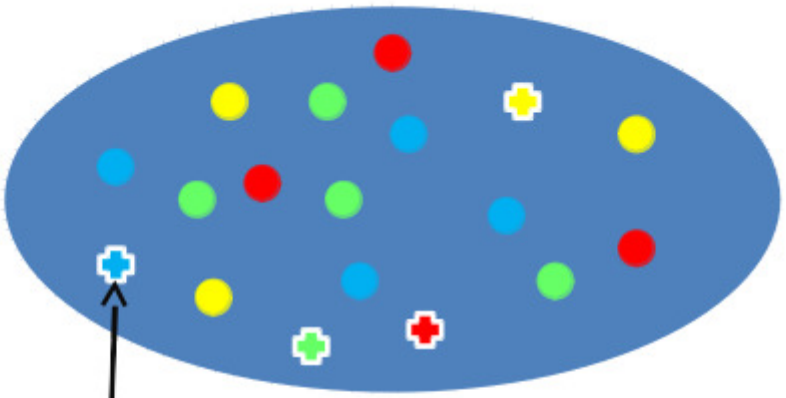
1. Let K be the number of clusters present in a chromosome
2. Centers of a string:
   $\{\overline{C}_1^d, \overline{C}_2^d, \ldots, \overline{C}_K^d\} = \{C_1^1, C_2^1, \ldots, C_d^1, \ldots, C_1^K, C_2^K, \ldots, C_d^K\}$;
3. For each view we first select the corresponding subset of cluster centers and repeat the following steps
   - A particular data point $\overline{x}$ is assigned to the cluster center $\overline{C}_j^v$ which is closest to it among all the cluster centers.

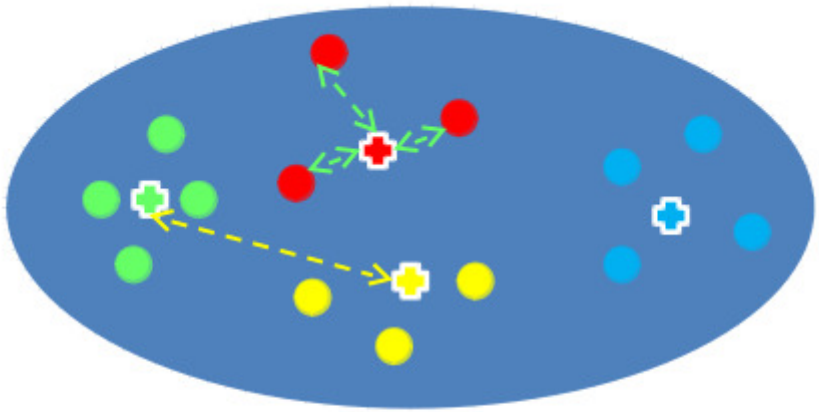
$$\pi_j^v = \{\forall \overline{x}_i^v \in \overline{X}^v : d(\overline{x}_i^v, \overline{C}_j^v) < d(\overline{x}_i^v, \overline{C}_l^v), l \neq j\} \tag{2}$$

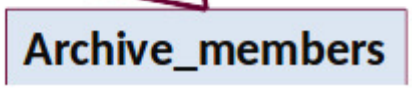
   - The new cluster centers are calculated as follows:

$$\overline{c}_j^v = \frac{\sum_{\overline{x}^v \in \pi_j^v} \overline{x}^v}{n_j^v} \tag{3}$$

Example of Point Assignment







	P_1	P_2	P_3	P_4	..	..	P_n
C_1	0	1	0	1	0	0	0
C_2	1	0	0	0	0	1	0
C_3	0	0	1	0	0	0	1
C_4	0	0	0	0	1	0	0

P_1 got assigned to cluster C_2

Objective Functions Used I

- PBM-index [2] is calculated for the final partitioning obtained using view v. Let the value be denoted by PBM_v.

$$PBM(K) = (\frac{1}{K} \times \frac{\mathcal{E}_1}{\mathcal{E}_\mathcal{K}} \times D_K) \tag{4}$$

  Here, K denotes the number of clusters, $\mathcal{E}_\mathcal{K} = \sum_{k=1}^{K} \sum_{j=1}^{n_k} d(\overline{c}_k, \overline{x}_j^k)$ and $D_K = \max_{i,j=1}^{K} d(\overline{c}_i, \overline{c}_j)$,

- A new objective function *Agreement Index (AI)* is calculated ; $n_a = \sum_{i=1}^{n} \sum_{j=1}^{n} I_{A_{ij}^{v1}, A_{ij}^{v2}}$, here

$$I_{A_{ij}^{v1}, A_{ij}^{v2}} = 1 \quad \text{if} \quad A_{ij}^{v1} = A_{ij}^{v2} \tag{5}$$

$$= 0 \quad \text{otherwise} \tag{6}$$

Objective Functions Used II

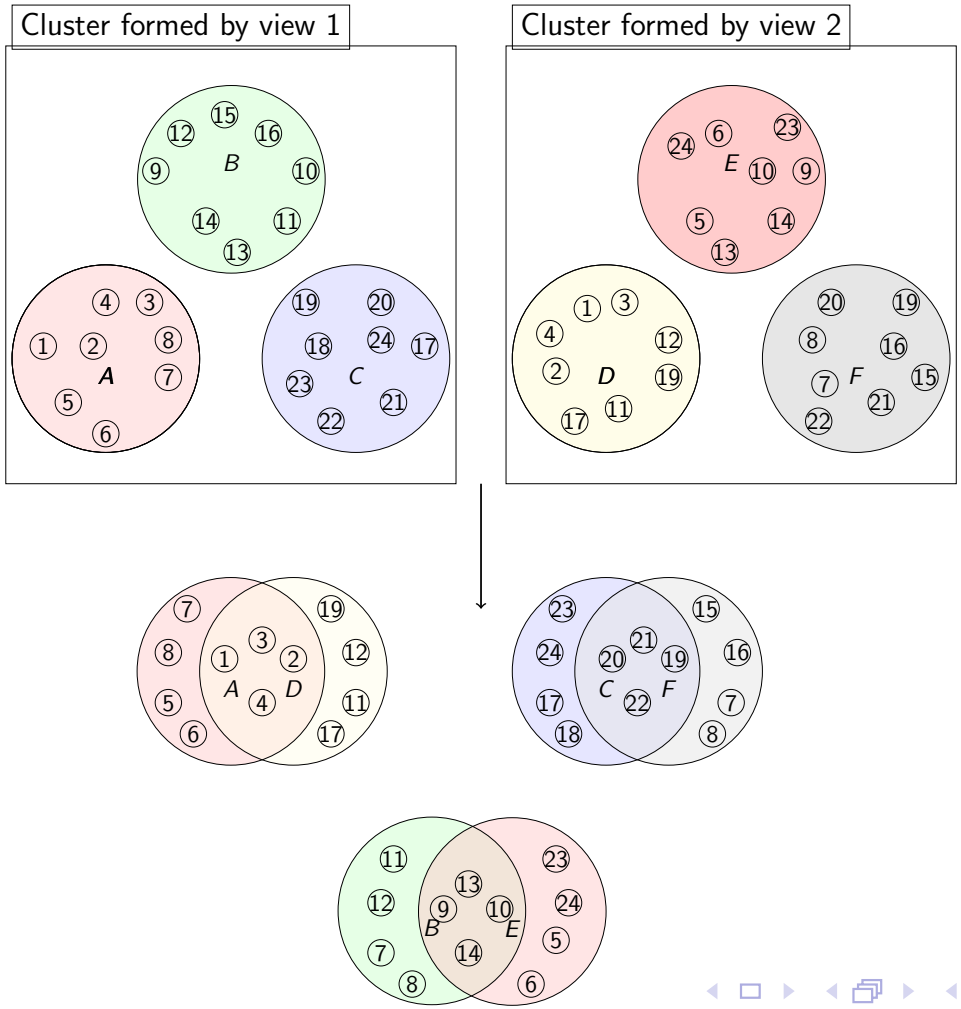
- $$\begin{aligned} A_{ij}^{v} &= 1 \text{ if } \overline{x}_i \text{ and } \overline{x}_j \text{ belong to the same cluster} \\ &= 0 \quad \text{otherwise} \qquad (7) \\ &= 1 \quad \text{if } i = j \qquad (8) \end{aligned}$$
- $n_d = n^2 - n_a$
- $$AI_{v1,v2} = \frac{n_a + 1}{n_d + 1} \qquad (9)$$
- $$AI = \frac{\sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} 2 \times AI_{v_i, v_j}}{m \times (m-1)}, \qquad (10)$$
- The objective functions corresponding to a particular string are:

$$\{PBM_1, \ldots, PBM_m, AI\}$$

Objective Functions Used III

- Steps of AMOSA are used to maximize all these objective functions simultaneously

Updation of Strings



Mutation Operation

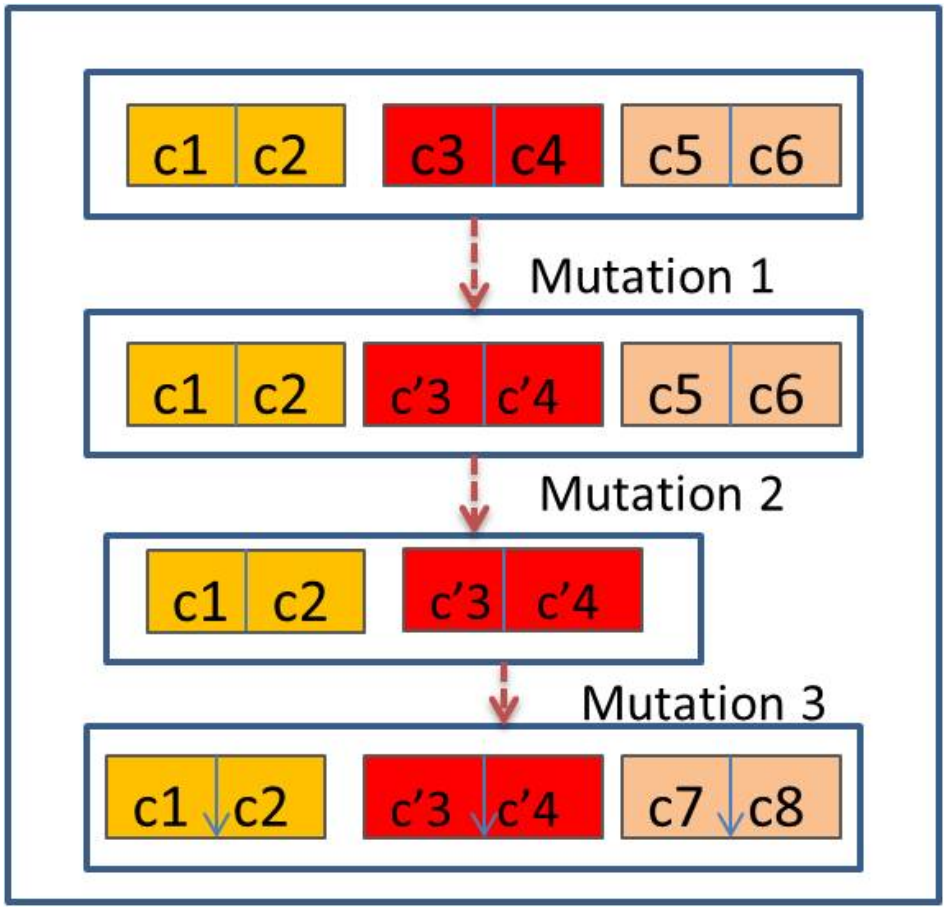


Figure: Example of mutation.

- Three types of mutation operations applied to generate new solution through AMOSA [4] process.
  - In Mutation 1, some cluster centres present in a state are modified by some value.
  - In Mutation 2, the total number of clusters present in a state is decreased by 1.
  - In Mutation 3, the total number of clusters present in a state is increased by 1.

Flowchart-I

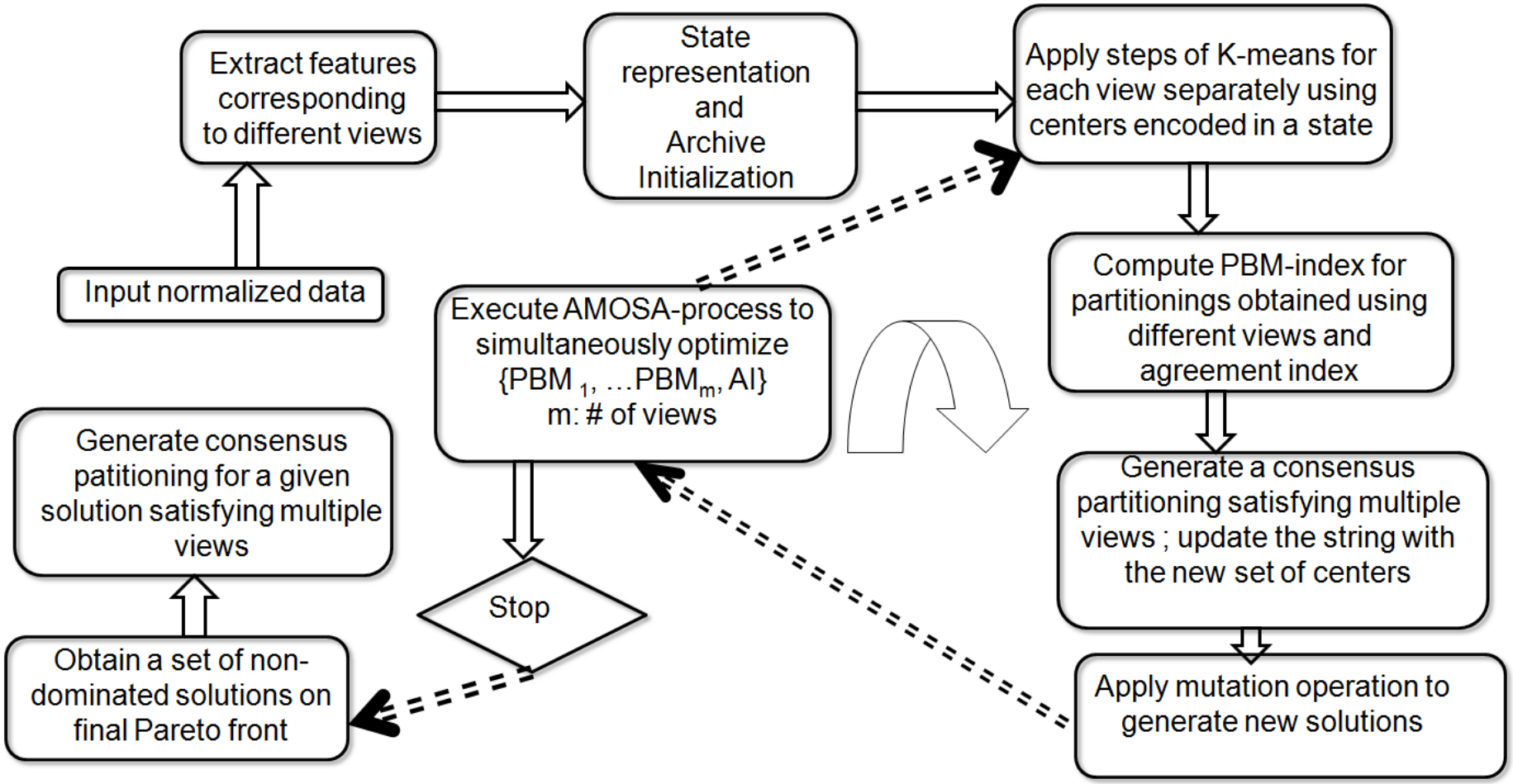


Figure: Flow chart of the first proposed approach (MOO-Multi1)

Flowchart-II

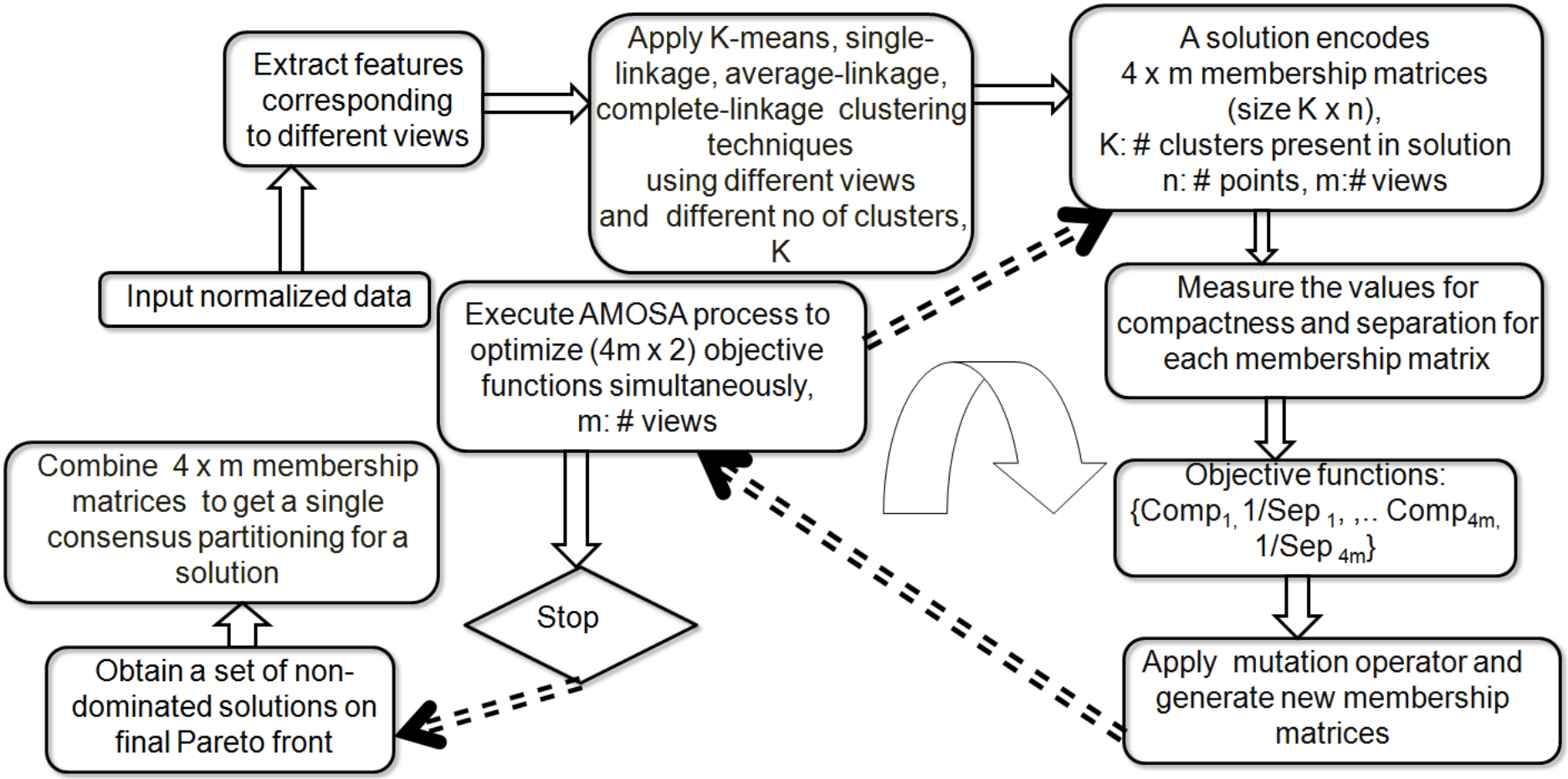


Figure: Flow chart of the second proposed approach (MOO-Multi2)

Flowchart-III

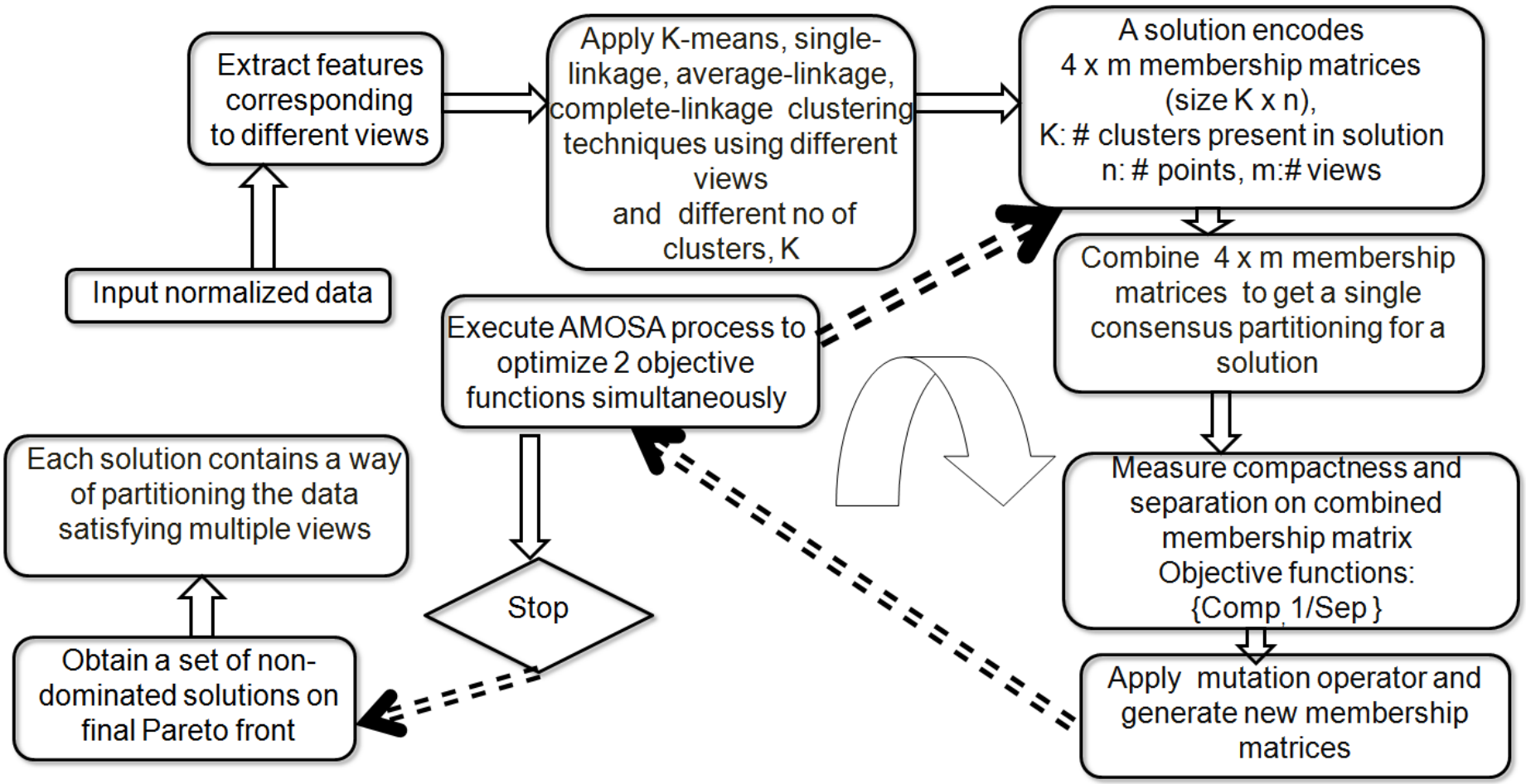


Figure: Flow chart of the third proposed approach (MOO-Multi3)

Experimental Results

- First batch of experiments: standard data from UCI Machine Learning Repository[1].
- Two views for each data set
- First view : original set of attributes
- Second view : apply Principal Component Analysis to the data set. The set of attributes captured using 95% of the data variance are used as second view.
- Results taken for three approaches proposed, multiview-KM [5], multi-objective spectral clustering [6]

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html

Description of UCI Datasets

Table: Description of Data Sets

Datasets	AC	Instances	Actual no. of features	Features used in View1	No. of PCA features/ Features used in view2	Total features
Iris	3	150	4	4	2	6
Newthyroid	3	215	5	5	2	7
Liver Disorder	2	345	7	6	4	10
Glass	6	214	10	9	4	13
Breast Cancer	2	699	10	9	5	14
Wine	3	178	13	13	6	19
Ionosphere	2	351	34	34	10	44
Leaf	40	340	16	14	7	21
Yeast	10	1484	8	8	4	12
Ecoli	8	336	8	7	4	11
Mice Protein Expression	8	1080	82	80	40	120
Diabetic Retinopathy Debrecen	2	1151	20	19	4	23
HTRU2	2	17898	9	8	4	12
Statlog (Shuttle)	7	58000	9	9	5	14
CovType	7	581012	54	54	26	80

Results of Mutual Information

Table: Results in terms of the maximum *Mutual Information* [7]. High values of *Mutual Information* [7] correspond to good partitionings. The best values for different data sets are marked in bold.

Dataset	MOO-Multi1$_{PBM}$	MOO-Multi$_{XB}$	MOO-Multi2	MOO-Multi3	multi-KM	MOO-Spectral
Iris	**0.7125**	0.6856	0.6792	0.0.6613	0.6598	0
Newthyroid	**0.7676**	0.7252	0.6984	0.5729	0.6759	0.0905
LiverDisorder	**0.3317**	0.3124	0.1591	0.1424	0.1233	0.0045
Glass	**0.4743**	0.4023	0.3152	0.2160	0.2753	0.2511
BreastCancer	**0.7399**	0.7234	0.6989	0.6601	0.6899	0.00087337
Wine	**0.8082**	0.79934	0.7944	0.7225	0.78544	0.4942
Ionosphere	**0.1411**	0.1331	0.1315	0.1281	0.1284	0.0797
Leaf	**0.67965**	0.37175	0.28476	0.3957	0.17595	0.1175489
Yeast	**0.23599**	0.22135	0.211354	0.210024	0.2080	0.059423693
Ecoli	0.60666	**0.63669**	0.58974	0.62864	0.56623	0.0119
Mice Protein Expression	**0.08482176**	0.0828745	0.07947312	0.0803421	0.0075948	0.071055
Diabetic	**0.1075100047**	0.00514921	0.0085784	0.0964849	0.0015973	0.0072
HTRU2	**0.2404177703**	0.2203145	0.13887	0.148957	0.154762	0.1441
Statlog(Shuttle)	**0.30174**	0.2796385	0.124967	0.133862	0.12687	0.1158
CovType	0.1906001	0.1607001	**0.2354921**	0.2054921	0.1307959	0.002500292

Results of Adjusted Rand Index

Table: Results in terms of the maximum *Adjusted Rand Index* [8]. High values of the *Adjusted Rand Index* [8] correspond to good partitioning. The best values for different data sets are marked in bold.

Dataset	MOO-Multi1$_{PBM}$	MOO-Multi$_{XB}$	MOO-Multi2	MOO-Multi3	multi-KM	MOO-Spectral
Iris	**0.67**	0.6561	0.635	0.588	0.6201	0
Newthyroid	**0.78**	0.7112	0.614	0.6014	0.5974	0.2453
LiverDisorder	**0.15**	0.1212	0.1057	0.1110	0.06104	−0.0082
Glass	**0.36**	0.310	0.15366	0.14148	0.16107	0.3467
BreastCancer	**0.99**	0.98	0.8538	0.8437	0.8337	−0.0083
Wine	**0.662679**	0.6423	0.62835	0.6048	0.6148	0.5270
Ionosphere	**0.28**	0.243	0.2031	0.2135	0.1679	−0.0014
Leaf	**0.3255584**	0.23628	0.248962	0.25896	0.0189547	0.115746
Yeast	**0.1261951**	0.121016	0.1129	0.10689	0.10474	0.10247
Ecoli	0.655776	**0.67409**	0.66409	0.6757	0.37425	0.2048
Mice Protein Expression	**0.0945000413**	0.0934587219	0.01968	0.0200985	0.062849	0.0219
Diabetic	**0.118363532**	0.099873612	0.01097845	0.0115874	0.010087	−0.0016
HTRU2	**0.3537545029**	0.334178	0.27415	0.29877	0.29475	0.1247
Statlog(Shuttle)	**0.4585491**	0.448914	0.3108954	0.3708954	0.3695123	0.3648
CovType	0.1169999	0.1069749	**0.1315447**	0.1217457	0.0494089	0.000001724352

Results of Minkowski Score

Table: Results on some real-life data sets; the minimum *Minkowski Score* [9] values obtained by different clustering algorithms are reported. Low values of the *Minkowski Score* [9] correspond to good partitionings. The best values for different data sets are marked in bold.

Dataset	MOO-Multi1$_{PBM}$	MOO-Multi$_{XB}$	MOO-Multi2	MOO-Multi3	multi-KM	MOO-Spectral
Iris	**0.6598**	0.6598	0.72881	0.8888	0.7142	0.99
Newthyroid	**0.4638**	0.5078	0.5376	0.5436	0.6089	0.5635
LiverDisorder	**0.9631**	0.9846	0.9988	0.969	0.998621	0.9778
Glass	1.09	1.1415	1.137	1.189	1.1459	**0.6439**
BreastCancer	**0.32**	0.33	0.3489	0.3562	0.3876	0.9076
Wine	**0.662679**	0.6875	0.6976	0.7154	0.9618	0.6764
Ionosphere	**0.8697**	0.897689	0.9042	0.9165	0.9288	0.9534
Leaf	**1.16438**	1.24451	1.265	1.22515	1.54726	1.2048965
Yeast	**1.10384**	1.127864	1.1555	1.18057	1.18766	1.19975
Ecoli	0.72356	**0.72302**	0.7354	0.7283	0.883639	0.8124
Mice Protein Expression	**1.1966907998**	1.205684	1.34652	1.25745	1.36478	0.9827
Diabetic	**0.9900981664**	0.992583	1.12874	1.02894	1.105689	0.9962
HTRU2	**0.6550780146**	0.689214	0.70964	0.70498	0.695487	0.7285
Statlog(Shuttle)	**0.6007216**	0.626147	0.68479	0.64854	0.648921	0.6895
CovType	**1.360363**	1.380736	1.595827	1.698327	1.907646	2.447489

Description of SRC Datasets

- Application to a real-world problem: search result clustering.
- Web search results clustering (SRC), also known as post-retrieval clustering or ephemeral clustering, has become popular in the past twenty years for helping the users in web browsing.
- The goal is to group the available web snippets into some meaningfully labeled clusters. The web snippets are retrieved from a search engine for a given query

Table: SRC gold standard datasets.

Datasets	# of queries	# of Subtopics Avg / Min / Max	# of Snippets
AMBIDENT	44	17.95/6/37	4400
ODP-239	239	10 / 10 / 10	25580
MORESQUE	114	6.7 / 2 / 38	11402

View Generation from SRC Data Sets I

1. The syntactic view: This view represents the syntactic information of a document given a particular query.
   i) The terms in the documents are first extracted.
   ii) A document-term matrix is created to be used as view1.
   iii) TF-IDF (a common weighting scheme) [10] values are used to fill this matrix.
2. The semantic view: This view captures the semantic contents of the web snippets :
   i) A content-based similarity measure, Symmetric Conditional Probability (SCP) [11] is used to calculate the similarity between two web snippets.
   ii) Given two word feature vectors d_i and d_j (consisting of some important words present in the web snippets), the corresponding

View Generation from SRC Data Sets II

similarity is computed by considering the similarity of their constituents:

$$S(d_i, d_j) = \frac{1}{\|d_i\|\|d_j\|} \sum_{r=1}^{\|d_i\|} \sum_{b=1}^{\|d_j\|} SCP(w_i^r, w_j^b) \tag{11}$$

$$SCP(W_1, W_2) = \frac{P(W_1, W_2)^2}{P(W_1) \times P(W_2)}, \tag{12}$$

Comparison of F1 measure of the proposed approach on SRC datasets.

Table: Evaluation results in terms of F 1 over MORESQUE and ODP239 data sets: Comparison of the proposed approach with state-of-the-art single view based approaches.

Datasets	MOO-multi1		MOO-clus		SOO-SRC			
	Min	Max	Min	Max	GK-means	STC	LINGO	BIK
MORESQUE	0.6689	**0.6987**	0.658	0.675	0.655	0.455	0.326	0.317
ODP-239	0.4059	**0.4294**	0.379	0.384	0.366	0.324	0.273	0.2

F1-Measure on Web clustering Search Results

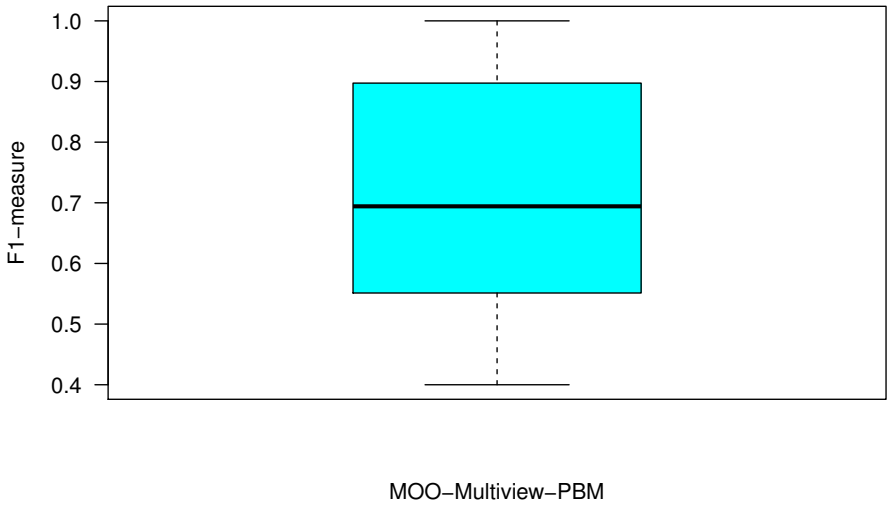


Figure: F1-measure values obtained for different queries combining all the data sets together after application of the proposed algorithm $MOO - Multi1_{PBM}$.

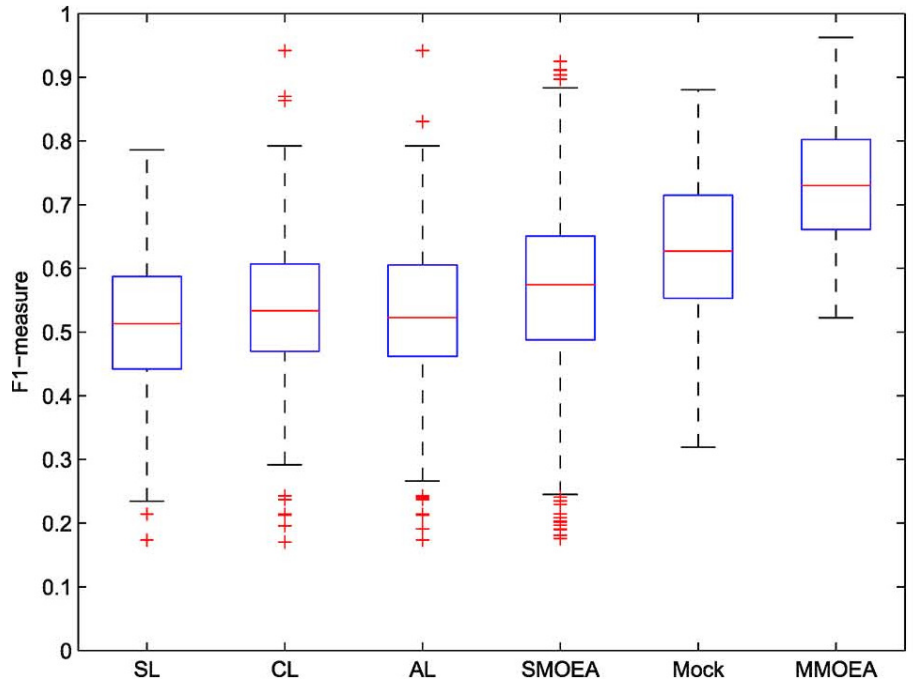


Figure: F1-measure values obtained for different queries combining all the data sets together after application of the algorithm MMOEA [12].

Precision on Web clustering Search Results

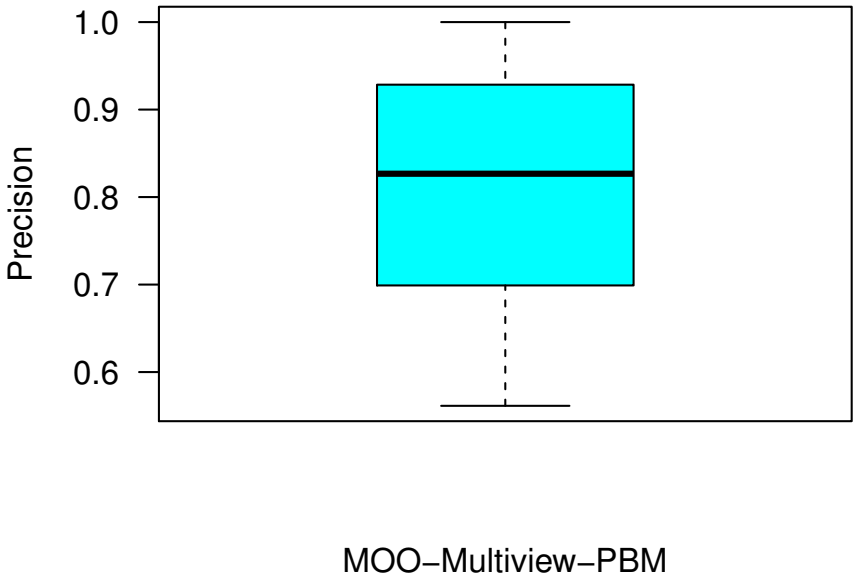


Figure: Precision values obtained for different queries combining all the data sets together after application of the proposed algorithm $MOO - Multi1_{PBM}$.

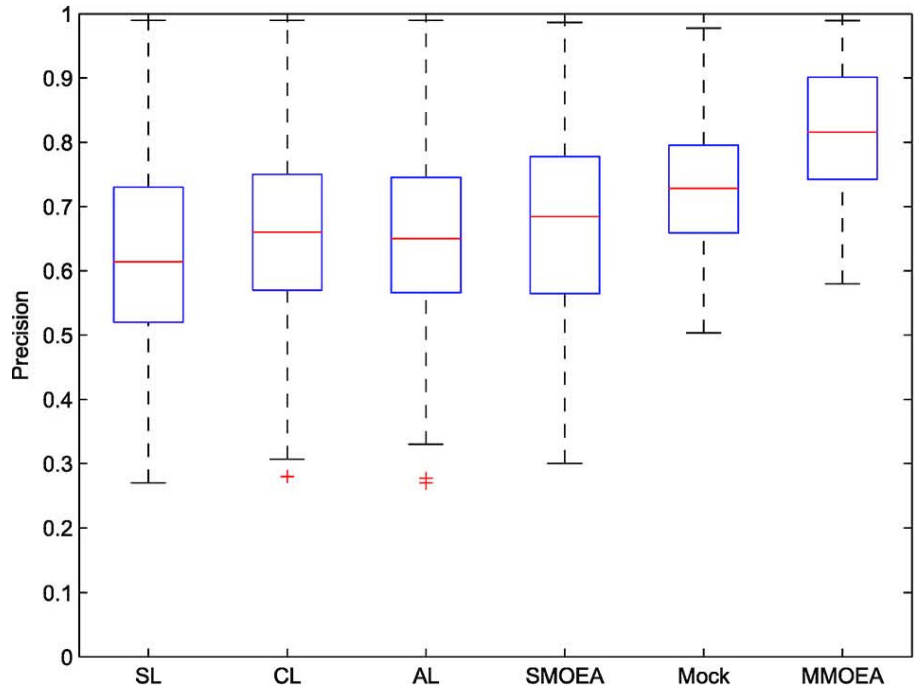


Figure: Precision values obtained for different queries combining all the data sets together after application of the algorithm MMOEA [12].

Recall on Web clustering Search Results

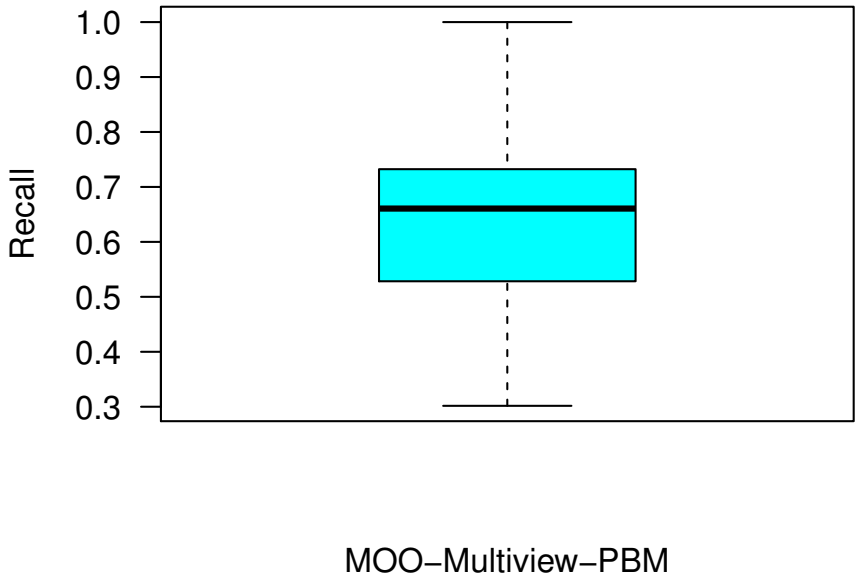


Figure: Recall values obtained for different queries combining all the data sets together after application of the proposed algorithm $MOO - Multi1_{PBM}$.

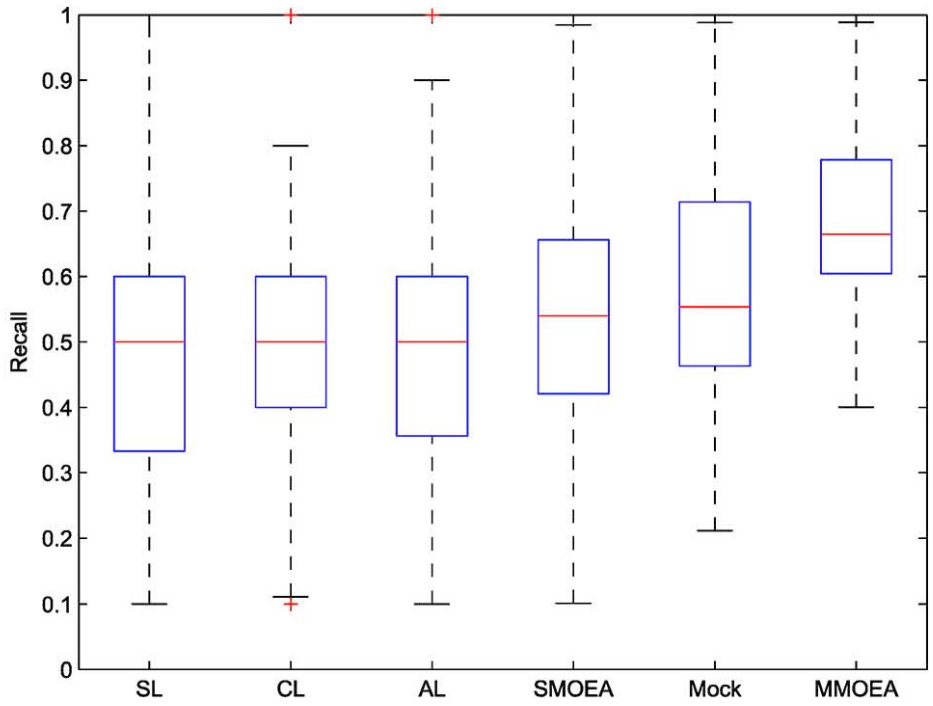


Figure: Recall values obtained for different queries combining all the data sets together after application of the algorithm MMOEA [12].

Rand-Index on Web clustering Search Results

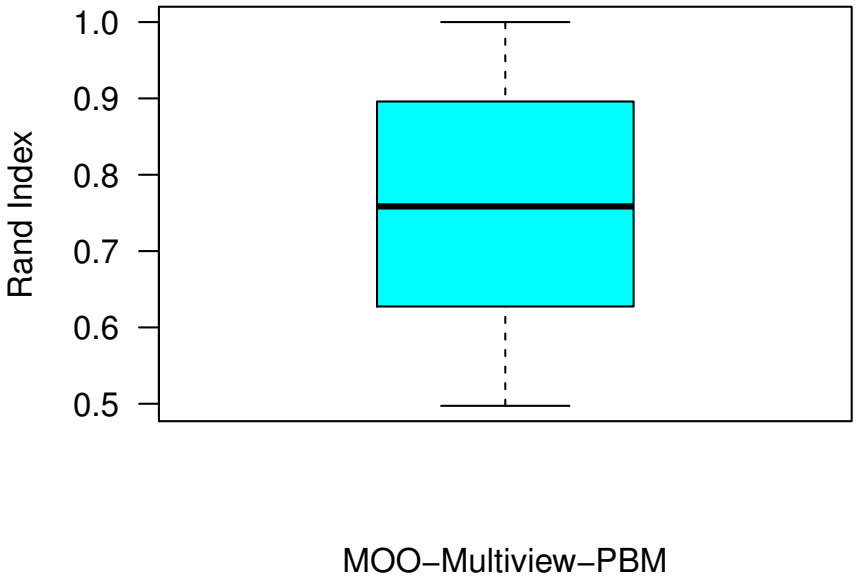


Figure: Rand Index values obtained for different queries combining all the data sets together after application of the proposed algorithm $MOO - Multi1_{PBM}$.

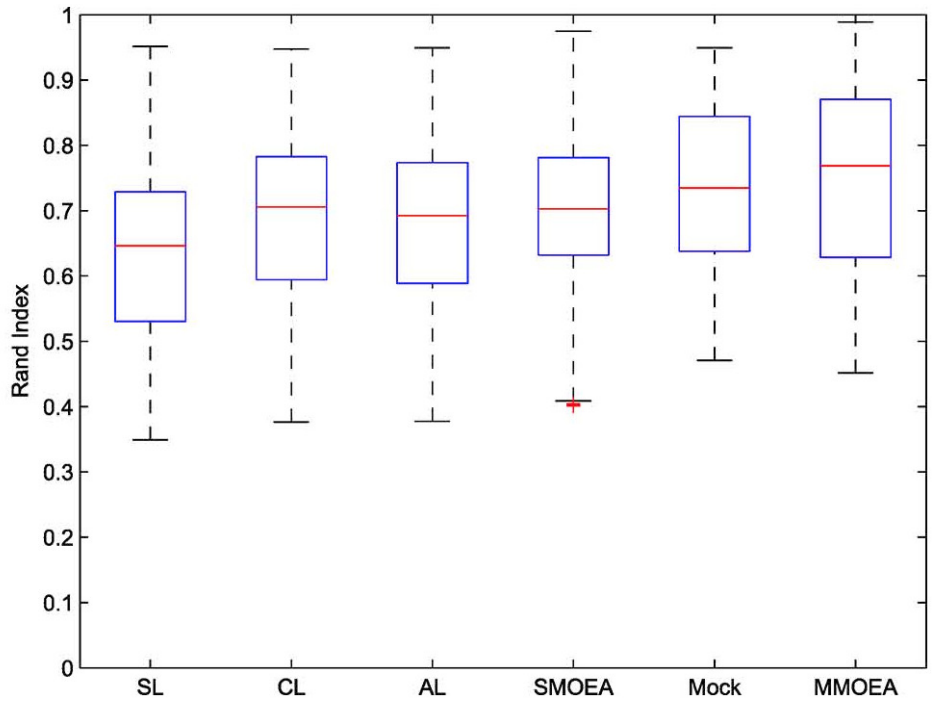


Figure: Rand Index values obtained for different queries combining all the data sets together after application of the algorithm MMOEA [12].

Conclusions and Future Works I

1. Some multiobjective multi-view clustering approaches are developed: i) partitional approach ii) cluster ensemble based approach.
2. Approaches can automatically determine the number of clusters present in a data set.
3. A single consensus partitioning satisfying multiple views is obtained.
4. *Agreement Index*: new quality measure to check the similarity between partitionings obtained using different views.
5. Results shown for UCI as well as search result clustering
6. MOO based partitioning approach provides better results than other methods.

Future works:

1. Use of some multiobjective algorithms in place of AMOSA to understand its capacity as the underlying optimization technique.

Conclusions and Future Works II

2. Automatically determining multiple views from a data set without using any domain knowledge
3. Apart from web search data, many other real-world domains like molecular biology (RNAseq gene expression and proteomics) can benefit from such techniques.

References I

- C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.
- M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition*, vol. 37, no. 3, pp. 487–501, 2004.
- X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 8, pp. 841–847, 1991.
- S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: Amosa," *IEEE transactions on evolutionary computation*, vol. 12, no. 3, pp. 269–283, 2008.
- S. Bickel and T. Scheffer, "Multi-view clustering." in *ICDM*, vol. 4, 2004, pp. 19–26.

References II

- X. Wang, B. Qian, J. Ye, and I. Davidson, "Multi-objective multi-view spectral clustering via pareto optimization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 234–242.
- L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.
- A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis," *Functional Genomics: Methods and Protocols*, pp. 159–182, 2003.
- S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.

References III

J. F. Da Silva, G. Dias, S. Guilloré, and J. G. P. Lopes, "Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units," in *Portuguese Conference on Artificial Intelligence*. Springer, 1999, pp. 113–132.

A. Wahid, X. Gao, and P. Andreae, "Multi-view clustering of web documents using multi-objective genetic algorithm." in *IEEE Congress on Evolutionary Computation*, 2014, pp. 2625–2632.

Thank You

Specials Thanks to Prof. Stefan Kramer for hosting me in his lab