Multi-Objective Clustering Ensemble for High-Dimensional Data based on Strength Pareto Evolutionary Algorithm (SPEA-II)

Abdul Wahid
Victoria University of Wellington
New Zealand
abdul.wahid@ecs.vuw.ac.nz

Xiaoying Gao
Victoria University of Wellington
New Zealand
xgao@ecs.vuw.ac.nz

Peter Andreae
Victoria University of Wellington
New Zealand
pondy@ecs.vuw.ac.nz

***Abstract*—Clustering is one of the fundamental data analysis techniques, which aims to find distinct groups of similar objects and discovers hidden structures in data. A recent clustering approach, *clustering ensembles* tries to derive an improved clustering solution based on previously generated different candidate clustering solutions.**

Clustering ensembles have two steps: *generating multiple candidate clustering solutions* from the data and *forming a final clustering solution* from previously generated candidate clustering solutions.

A problem of the first step is the text representation, where word frequencies are often used as features. Other semantic information of the text such as topics, hypertext, etc are ignored. The problem for the second step is that the current popular median partition approach selects one clustering solution from previously generated candidate clustering solutions.

A common clustering ensemble approach uses word frequencies as features to represent text data (documents). However, documents usually contain semantically rich information i.e. words, hypertext, titles, topics etc. The cluster ensemble approach ignores the semantic information of the documents and hence is prone to produce futile groupings of the documents.

In this research work, we present a new multi-objective clustering ensemble method based on Strength Pareto Evolutionary Algorithm (SPEA-II).

Our method utilizes the semantic information (rich features) to address the first problem of clustering ensembles. The cluster oriented evolutionary approach which derives the final clustering solution by selecting better quality clusters is in the second step of our method to address the second problem. The results show that our new method provides better results than other clustering ensemble methods.

***Keywords*—Clustering Ensemble, Multi-Objective Optimization, Evolutionary Algorithm**

I. Introduction

Clustering is a popular technique which groups data objects into different clusters. There are many clustering methods for clustering different types of the data. Every clustering method produces a clustering solution and its very likely to have different clustering solutions of same datasets. In recent years, a clustering approach *Clustering Ensembles* is getting much attention because it produces fairly better results by combining the results of different clustering methods [1].

Clustering ensemble methods consist of two steps: in first step candidate clustering solutions are generated, and in the second step a single candidate clustering solution is derived from previously generated candidate clustering solutions.

Median partition based clustering ensembles, which selects a single candidate clustering solution from a set of candidate clustering solutions is so far considered as the best approach [2]. Selecting a single candidate clustering solution is commonly based on a similarity criterion. This similarity criterion picks a candidate clustering solution that has a maximum average similarity to all previously generated candidate clustering solutions. This approach assumes that the first step of the clustering process will generate a similar clustering solution, however, the results generated from multiple clustering algorithms might differ from each other.

Generally, clustering ensemble methods formulate clustering as a single objective optimization problem. However according to recent studies, using two or more objectives leads to better results [3], [4], [5]. In order to solve multi-objective clustering problem, multi-objective evolutionary algorithms such as SPEA-II and NSGA-II are widely applied in clustering ensemble. Evolutionary algorithms heavily depend upon evolutionary operators (such as selection, crossover, mutation etc) and fitness evaluation criteria. The evolutionary operators and the evaluation criteria are domain specific. The current multi-objective clustering ensembles based on evolutionary algorithms use simple crossover and mutation methods along with fitness evaluation that do not penalize a clustering solution having small clusters.

Moreover, as different clustering solutions generally include high and low quality[1] clusters, therefore a final clustering solution selected by partition based approach might not be the best clustering solution.

This research work presents a new multi-objective clustering ensembles method (*MDC*) based on evolutionary approach and uses multiple views to generate a diverse set of candidate clustering solutions in the initial step. The concept of multiple views in documents means representing a document in more

978-1-4673-8273-1/15/$31.00

[1]The quality of cluster is generally measured by computing the intra-cluster and inter-cluster distances.

than one way e.g. using term frequencies, topics, hyperlinks etc to represent the documents. Hence, the MDC not only uses the word frequencies but also uses the semantic information of the text.

MDC also addresses the limitations of current multi-objective clustering ensembles methods that are based on evolutionary algorithms, by developing new evolutionary operators, better fitness evaluation functions and a cluster oriented approach, which forms the final clustering solution by selecting high quality clusters from different clustering solutions, rather than selecting one candidate clustering solution. Key contribution of this research work as are follows.

1) Introducing the concept of multiple views in the initial step of clustering process for generating diverse candidate clustering solutions.
2) Developing crossover methods for generating a new set of clusters from previous candidate clustering solutions.
3) Developing guided mutation methods for splitting and merging clusters in a candidate clustering solution.
4) Developing multi-objective function based on intra-cluster and inter-cluster distances for multi-objective optimization.

The rest of the paper is organized as follows. Section II discusses related work, section III describes our method (MDC), section IV provides the details of the experimental setup and the results, section V discusses the analysis of MDC and section VI concludes the paper.

II. Background and Related Work

Most of the clustering methods provide partitioning of the data and do not allow overlapping within the clusters. Moreover, current clustering ensemble methods only use a single view of the data and mainly focus on selecting the one clustering solution from multiple candidate clustering solutions. Also, some of the clustering methods implement clustering as a single objective optimization problem. The multi-objective approaches for clustering methods are as follows.

Korkmaz et al. proposed a multi-objective approach introducing two objectives and used linkage based encoding scheme to reduce the redundancy of the initial clustering solutions [6]. Their first objective minimizes the number of clusters and the second minimizes the intra-cluster variance. This approach utilized the concept of Pareto dominance to find a set of nondominated clustering solutions.

Bandyopadhyay et al. proposed a fuzzy clustering method using multi-objective evolutionary algorithm [7]. Their approach also used two objective functions: J_m criterion [8] and Xie-Beni index [9].

A closely related method to our research work is MOCK, a cluster ensemble method proposed by Handl and Knowles [5], [10], which implements two objectives. Their method used a modified version of SPEA-II algorithm and used two objective functions: connectedness and compactness of the cluster. These two objectives were inspired from the single link and the k-means algorithm.

Other popular approaches for multi-objective clustering ensembles algorithms are [11], [12]. In general, multi-objective clustering ensembles proposed so far use two objectives and they mainly focus on minimizing the intra-cluster variance. A comprehensive analysis of multi-objective approaches can be found in [13].

Apart from clustering ensemble methods, the use of multiple views in the clustering process is mainly referred as multi-view clustering [14], [15]. Our new method, MDC is related to NSGA-II based multi-objective clustering method (MMOEA) [16], [17], which is a clustering ensemble method that uses multiple views and implements three objective functions. The results comparison of MOCK, MMOEA and our new method is provided in the results section.

III. The Method

Our Multi-objective Document Clustering (MDC) method is based on clustering ensemble approach using SPEA-II [18]. It exploits multiple views of the documents to generate various candidate clustering solutions in step 1 of the clustering process and then forms a final clustering solution by combining a set of high quality clusters from different candidate clustering solutions using evolutionary process. The method uses two criteria: intra-cluster distances and inter-cluster distances to evaluate the fitness of individuals and performs Selection, Crossover, Mutation and Tuning steps on the population to generate a new set of candidates.

A. Initial Population

Initial population i.e. the initial set of candidate clustering solutions of documents are generated using two views: content and hypertext of the web pages[1]. We constructed two feature matrices using the standard Term Frequency Inverse Document Frequency (TFIDF) of the content and Term Frequency (TF) of the hypertext of the web pages. Then we applied traditional clustering algorithms: single link, complete link, average link (hierarchical methods) and k-means with different initialization on two feature matrices separately. This resulted in eight different candidate clustering solutions.

B. Genetic Representation

We chose a matrix based binary encoding scheme [19] to represent the individuals. This enabled us to have overlapping clusters. Figure 1 depicts an individual clustering solution and its matrix based encoding is shown in Figure 2. The rows represents clusters and columns represents the document. The value 1 means the document is assigned to the cluster.

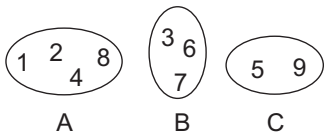

Fig. 1: Example of a Clustering Solution.

[1]These two views content and hypertext were predefined and provided with the WEBKB dataset.

	1	2	3	4	5	6	7	8	9
A	1	1	0	1	0	0	0	1	0
B	0	0	1	0	0	1	1	0	0
C	0	0	0	0	1	0	0	0	1

Fig. 2: Sample individual/ clustering solution encoded as a binary matrix. The rows represent clusters (A, B and C) and columns represent the documents (1-9)

C. Fitness Evaluation

We considered two criteria for evaluating the fitness of an individual. The first criterion is intra-cluster distance penalized by the inverse square root of number of objects in a cluster and the second criterion is the inter-cluster similarity. These two criteria needs to be minimized and hence they were implemented as two objective function that needs to be minimized. The first objective functions is defined as follows:

$$Obj_1(C) = \frac{1}{|C|}\sum_{c\in C}\frac{1}{\sqrt{|c|}}\left(\frac{1}{|c|(|c|-1)}\sum_{d,d'\in c}\delta(d,d')\right)$$

where c is a cluster in clustering solution C, d and d' are documents in a cluster c, δ is a function which computes the distance between two documents. $\gamma(d,d')$ represents cosine similarity measure (commonly used in text clustering) of documents d and d' whereas $\delta(d,d')$ represents the 1-γ. The cosine similarity measure is widely used in finding the similarity between documents [20]. Following is the definition of the second objective function:

$$Obj_2(C) = \frac{1}{|C|(|C|-1)}\sum_{c,c'\in C\wedge c\neq c'}\left(\frac{1}{|c||c'|}\sum_{d\in c,d'\in c'}\gamma(d,d')\right)$$

where c is a cluster in clustering solution C, d and d' are the document in two different clusters c and c'. γ is the cosine similarity function which computes the similarity between two documents.

We implemented the objective functions in such a way that it becomes a multi-objective optimization problem where both objectives Obj_1 and Obj_2 are required to be minimized.

D. Selection

The selection step in our evolutionary process selects a number of individuals from the previous population to generate individuals. We used binary tournament method to select the individuals.

E. Crossover

After selecting two parents, we performed the crossover steps. This research work uses two crossover methods: the row-wise and the column-wise crossover method.

Figure 3 shows the row-wise crossover method. Parent 1 and Parent 2 randomly exchange two rows (dark gray background) and results in two new children (Child 1 and Child 2). The light gray row in Child 1 (cluster D) indicates unassigned documents after the crossover. Hence, this method produces new clusters to collect leftover documents. This crossover method can also produce overlapping clusters (a document is assigned to multiple clusters) and the coverage (assignment of all documents to clusters) is always 100%.

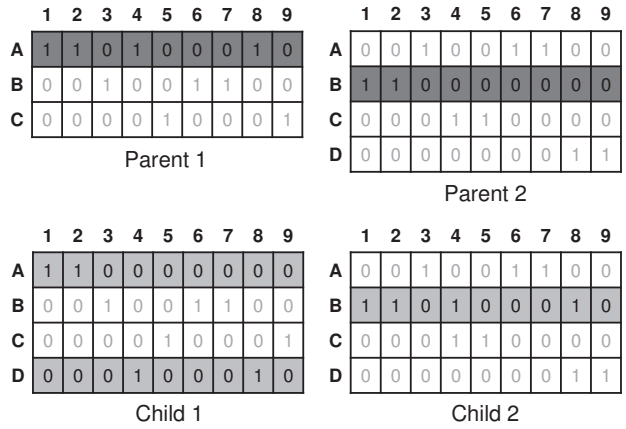

Fig. 3: Row-wise crossover method.

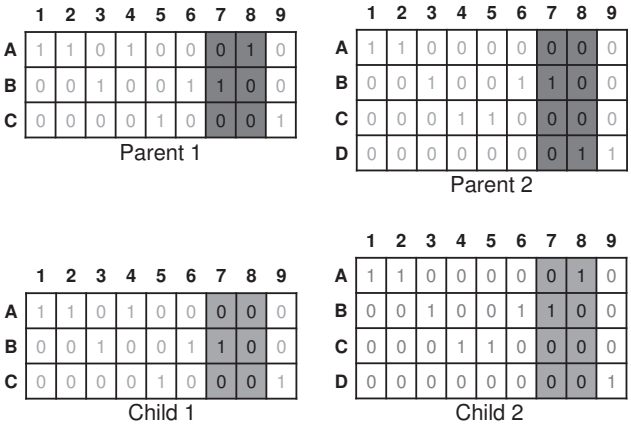

Fig. 4: Column-wise crossover method.

Figure 4 shows the column-wise crossover method. Parent 1 and Parent 2 exchange corresponding columns (with dark gray background) and results in two new children (Child 1 and Child 2). Note that the Child 1 after crossover contains zero value for column 8, indicating that the document is not assigned to any cluster. Since this crossover can produce empty clusters, we perform an additional step to remove any empty cluster. The coverage in this crossover method can be less than 100% (i.e. not all documents are assigned to clusters).

F. Mutation

We developed two types of mutation methods: split-mutation and merge-mutation. The split-mutation, splits one big cluster into two small clusters and the merge-mutation, merges two small, but similar clusters into one big cluster. The splitting of the cluster is based on random approach whereas

the merge-mutation is performed by considering inter-cluster distances of the two clusters.

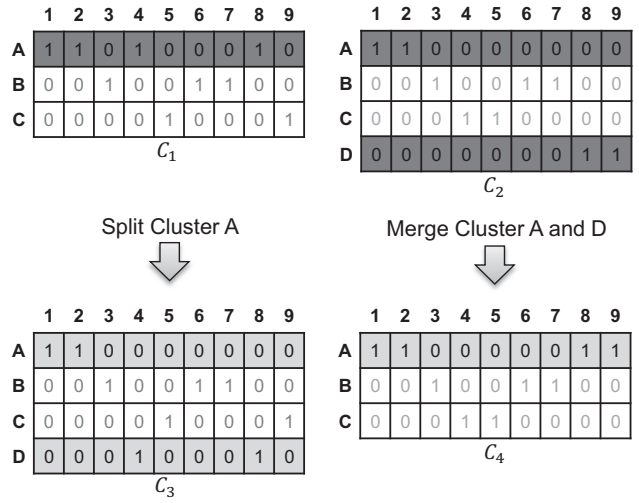

Fig. 5: Example of split and merge mutation method.

Figure 5 shows examples of split and merge mutation methods. The split-mutation is shown on left where cluster A of C_1 clustering solution is randomly divided into cluster A and D. The merge-mutation is shown on the right where cluster A and D in C_2 are merged into cluster A.

G. Tuning

We performed a tuning step after generating the new individuals to refine the clusters and help the evolutionary process to converge quickly. This tuning step is similar to the second step of the k-means algorithm. We iteratively calculate the cluster centroids and relocate the documents to a cluster based on the minimum distance of the document to the cluster centroid. The tuning process is repeated until the centroids remained the same and no document was required to be relocated. The centroid of a cluster is computed using the average similarity of all the documents in that cluster. Our implementation does not tune the overlapping clusters, hence overlapping between clusters are preserved.

H. MDC Algorithm

Algorithm 1 shows the process of our new method MDC. The algorithm is a modified version of SPEA-II and starts with the initial population, which is generated by exploiting the multiple views of the documents. Then it computes the objective functions Obj_1 and Obj_2 for each individual in a population. Then using the SPEA-II method, we identify the non-dominated solutions and finally perform our selection, crossover, mutation and tuning steps (as described earlier) on the current population to generate a new population. The truncation operator, mentioned in the algorithm, removes 50% of the worst individuals in a population by considering their fitness value. The stopping criterion for the algorithm is the total number of generations.

Algorithm 1 MDC - Document Clustering based on SPEA-II

1: Generate and represent encoded initial population Q_0 from different views of the documents. Set archive $P_0 = 0$. Set population size N. Set archive size N_p and Set generation $g = 0$.
2: Compute the objective functions Obj_1 and Obj_2 of all individuals in Q_g and P_g
3: Copy all non-dominated solutions in $Q_g + P_g$ to P_{g+1}
4: **if** $|P_{g+1}| > N_p$ **then**
5: Reduce P_{g+1} by means of truncation operator
6: **else if** $|P_{g+1}| <= N_p$ **then**
7: Copy N_p - $|P_{g+1}|$ dominated solutions from $Q_g + P_g$ to P_{g+1}
8: **end if**
9: **if** Stopping criterion is satisfied **then**
10: return P_{g+1}
11: **end if**
12: **for** $i \leftarrow 1, N/2$ **do**
13: $p_1, p_2 \leftarrow$ select two individuals from P_{g+1}
14: $rand \leftarrow$ generate a random number from 1 to 6.
15: **if** $rand = 1$ **then**
16: $c_1, c_2 \leftarrow$ perform row-wise crossover(p_1, p_2)
17: **else if** $rand = 2$ **then**
18: $c_1, c_2 \leftarrow$ perform column-wise Crossover(p_1, p_2)
19: **else if** $rand = 3$ `or` $rand = 4$ **then**
20: $c_1, c_2 \leftarrow$ perform merge-mutation if possible, otherwise apply split-mutation on p_1 and p_2
21: **else if** $rand = 5$ `or` $rand = 6$ **then**
22: $c_1, c_2 \leftarrow$ perform split-mutation if possible, otherwise apply merge-mutation on p_1 and p_2
23: **end if**
24: $Q_{g+1} \leftarrow$ Apply tuning on c_1 and c_2
25: **end for**
26: Go to 2

IV. Experiments Setup and Results

The algorithm was run multiple times and the average of all the runs are reported in this research work. The parameters for the algorithms include: maximum number of generations was set to 1000, crossover probability was set to $\frac{1}{6}$, mutation probability was set to $\frac{1}{3}$, population size was set to 20. These parameters were fixed for all runs of the algorithm. The number of clusters were randomly chosen at initial stage ranging from 2-10 for k-mean clustering and fixed number of clusters (predefined number of clusters from gold standard) for other clustering methods.

MDC was compared with a simple ensemble clustering method (using an average link hierarchical method as consensus function) and two multi-objective clustering approaches based on evolutionary approaches (Mock and MMOEA). The results were compared using three different evaluation metrics on different datasets. These evaluation metrics are widely used to measure the quality of clustering by comparing the clusterings solutions produced by clustering methods with given clustering solution provided by the datasets[1].

[1]It is important to note that clustering is an unsupervised learning method and the evaluation metrics can not be used as objective functions because the evaluation metrics requires the gold standard clustering solutions, which are not present during the clustering process.

A. Datasets

We created eight datasets (D1-D8) from WebKB2 and WebKB4 datasets which have two predefined views of webpages[1]. Also, we used Citeseer [2] and Cora[3] dataset which also has two predefined views. Apart from two view datasets we used three view dataset which is combination of AMBIENT[4], MORESQUE[5] and ODP-239[6] datasets. The three views of combined dataset were generated from topics, terms and ambiguous queries of the document using wikiminer toolkit [21]. The following are the details of the datasets

1) WebKB2 dataset contains 1051 web pages from Cora, Texas, Washington and Wisconsin universities. We divided the dataset according to universities and constructed four datasets WebKB2-Cornell, WebKB2-Texas, WebKB2-Washington and WebKB2-Wisconsin. These four datasets (D1-D4) were then pre-processed by applying tokenization and lemmatization on each web page and two views were generated using Term Frequency Inverse Document Frequency (TFIDF) technique on terms of the web page and Term Frequency (TF) on Hypertext of the inlinks of that web page. The WebKB2 dataset labeled the web pages under two categories, course and non-course.
2) WebKB4 dataset contains 887 web pages from the Cora, Texas, Washington and Wisconsin universities. The web pages were clustered into course, faculty, student, project and staff categories according to the gold standard provided by the author of the dataset. After pre-processing, we constructed two views: a binary vector of terms, which specified if the term was present in the document, and Term Frequency of the Hypertext of the inlinks of web pages.
3) Citeseer dataset contained 3312 articles from citeseer publication database. These articles were categorized into six groups (Agents, AI, DB, IR, ML, HCI). After pre-processing, we constructed two views: a binary vector of terms and Term Frequency of the citations.
4) Cora dataset contained 2708 articles related to machine learning. These articles were categorized into seven groups (Case Base, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning and Theory). After pre-processing, we constructed two views: a binary vector of terms and Term Frequency of the citations.
5) Combined dataset contained 3970 documents which were retrieved against 397 queries. Each query had corresponding 100 short documents that were required to be clustered into different groups. The views were topics from documents, TFIDF of terms in the document and senses of the query. These three views were generated by the method specified in [16].

[1]Both of the datasets can be downloaded from http://www.cs.cmu.edu/~webkb/

[2]can be downloaded from http://www.cs.umd.edu/~sen/lbc-proj/data/citeseer.tgz

[3]can be downloaded from http://www.cs.umd.edu/~sen/lbc-proj/data/cora.tgz

[4]can be downloaded from http://credo.fub.it/ambient/

[5]can be downloaded from http://lcl.uniroma1.it/moresque/

[6]can be downloaded from http://credo.fub.it/odp239/

B. Validation Measure

The clustering performance was evaluated using the Clustering Accuracy (CA) [22], F1-measure (F1) [15], [23] and Rand Index (RI) [24]. F1-measure is defined as the weighted harmonic mean of Precision and Recall. Precision measures the accuracy of a system by considering the majority topics of the clusters and Recall measures the coverage of different topics in a clustering solution by considering the majority topics of the clusters [23]. The measures are defined as follows:

1) F1-measure is defined as the weighted harmonic mean of the Precision and Recall. Precision is the measure of the accuracy of a system whereas Recall is a measure for coverage of different topics in a clustering solution [23].
The precision or cluster accuracy of a cluster $C_i \in \mathcal{C}$ can be computed as:

$$P(C_i) = \frac{|C_i^t|}{|C_i|} \tag{1}$$

where C_i^t is the set of all the documents in cluster C_i which belong to the subtopic t[7]. The subtopic t is a majority subtopic shared by documents in cluster C_i. $|C_i|$ denotes the total number of documents in a cluster. The recall of a cluster that has a majority subtopic t is computed as

$$R(t) = \frac{|\bigcup C_{i \in \mathcal{C}^t} C_i^T|}{n_t} \tag{2}$$

where C^t is a subset of $\mathcal{C}$ whose subtopic t is in the majority. The n_t is the number of documents that belong to subtopic t. The total precision of the clustering solution and recall can be computed as

$$P = \frac{\sum_{C_i \in \mathcal{C}} P(C_i)|C_i|}{\sum_{C_i \in \mathcal{C}} |C_i|} \tag{3}$$

$$R = \frac{\sum_{t \in T} R(t) n_t}{\sum_{t \in T} n_t} \tag{4}$$

where T is a set of subtopics (gold standard).
2) Rand Index is widely used in literature to examine the agreement of newly developed clustering method with the ground truth (gold standard). The RI can be computed as:

$$RI(\mathcal{C}, \mathcal{G}) = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

where $\mathcal{C}$ is a clustering solutions and $\mathcal{G}$ is a ground truth. TP, TN, FP and FN are total number of true positives (pairs that are in the same cluster in a newly developed clustering solution and gold standard clustering solution), true negatives (pairs of documents that are in different clusters in both clustering solutions), false positive (pairs of documents that are in different clusters in $\mathcal{G}$ but in the same cluster in $\mathcal{C}$) and false negatives (pairs of documents that are in different clusters in $\mathcal{C}$ but in the same cluster in $\mathcal{G}$) respectively.

[7]the subtopic are provided by the datasets. they are considered as a gold standard

TABLE I: Clustering Accuracy computed on 10 different datasets.

Dataset	Avg-Ensemble		MOCK		MMOEA			MDC		
	v1	v1+v2	v1	v1+v2	v1	v1+v2	2view	v1	v1+v2	2view
D1	83.95	83.95	84.34	84.43	**91.12**	**91.12**	91.12	**91.12**	**91.12**	**92.51**
D2	86.22	87.94	87.01	87.39	**90.12**	90.21	90.42	**90.12**	**90.21**	**91.45**
D3	71.09	71.09	80.12	81.23	84.91	86.43	86.72	**87.05**	**88.12**	**92.13**
D4	73.83	74.16	74.98	75.33	85.54	86.96	88.54	**88.17**	**90.15**	**92.88**
D5	56.41	57.69	58.01	59.98	68.12	68.82	70.12	**70.12**	**71.13**	**75.22**
D6	72.36	60.17	72.36	60.17	73.01	74.87	75.45	**74.11**	**76.85**	**78.13**
D7	69.13	69.65	69.82	70.23	72.17	72.51	74.17	**72.31**	**72.54**	**77.93**
D8	78.11	78.49	**79.91**	80.45	**79.91**	80.98	79.91	**79.91**	**81.44**	**83.91**
Citeseer	43.87	44.32	45.98	45.98	50.01	51.45	54.34	**51.12**	**53.32**	**60.12**
Cora	44.72	46.35	46.51	47.22	50.95	51.12	55.11	**51.31**	**52.32**	**60.15**

TABLE II: F1-measure computed on 10 different datasets

Dataset	Avg-Ensemble		MOCK		MMOEA			MDC		
	v1	v1+v2	v1	v1+v2	v1	v1+v2	2view	v1	v1+v2	2view
D1	91.28	91.28	91.54	**91.68**	**91.95**	**91.68**	**92.01**	**91.95**	**91.68**	**92.01**
D2	**92.61**	92.61	92.56	**92.69**	**92.61**	**92.69**	**93.28**	**92.61**	**92.69**	**93.28**
D3	88.71	88.71	89.91	89.91	90.71	91.74	92.14	**91.03**	**92.11**	**93.82**
D4	84.38	86.45	86.12	86.93	86.32	87.23	89.91	**86.99**	**89.18**	**92.78**
D5	66.52	67.56	67.56	68.32	69.12	70.34	71.12	**70.12**	**71.61**	**75.82**
D6	65.09	47.97	66.01	48.13	67.19	67.73	70.01	**68.21**	**70.72**	**74.05**
D7	56.24	63.67	58.23	64.39	62.99	67.16	71.25	**63.87**	**69.63**	**74.42**
D8	67.87	68.01	70.03	70.97	74.32	75.46	76.23	**75.12**	**76.74**	**78.13**
Citeseer	52.87	56.78	54.01	58.94	58.87	59.73	61.23	**61.18**	**61.73**	**68.29**
Cora	55.54	56.53	57.98	58.17	59.62	61.15	62.21	**60.54**	**62.17**	**63.54**

TABLE III: Rand Index computed on 10 different datasets

Dataset	Avg-Ensemble		MOCK		MMOEA			MDC		
	v1	v1+v2	v1	v1+v2	v1	v1+v2	2view	v1	v1+v2	2view
D1	71.94	72.94	71.01	72.94	72.66	72.94	75.21	**73.13**	**73.64**	**75.43**
D2	76.14	76.14	76.99	76.34	77.89	78.31	79.89	**78.23**	**78.51**	**80.21**
D3	57.01	58.41	68.19	69.12	70.41	75.12	79.32	**71.11**	**78.12**	**87.12**
D4	59.38	61.55	69.11	70.33	74.93	74.93	75.58	**75.45**	**78.91**	**88.21**
D5	61.02	61.96	63.12	63.76	70.79	75.12	76.17	**73.12**	**75.97**	**82.72**
D6	66.73	48.13	67.53	50.11	72.65	73.12	74.65	**74.95**	**76.32**	**80.88**
D7	65.52	67.96	72.12	73.01	73.13	74.76	76.21	**74.41**	**75.11**	**80.56**
D8	75.58	76.83	75.88	77.63	80.76	81.12	82.12	**82.92**	**83.41**	**85.14**
Citeseer	60.43	73.06	72.23	74.34	76.26	77.13	77.23	**78.26**	**79.36**	**85.21**
Cora	70.21	73.92	76.12	74.23	78.79	80.21	81.83	**80.11**	**81.32**	**89.75**

C. Comparison on Two View Datasets

Our method MDC was compared with three clustering methods: AVG-Ensemble, MOCK and MMOEA. The first method is a single objective average-link clustering ensemble method based on link pairwise similarity matrices [25]. The second method MOCK is a multi-objective evolutionary algorithm [5] based on SPEA-II[1]. Our recent work, MMOEA is a multi-objective multi-view evolutionary algorithm based on NSGA-II approach [**?**]. MMOEA uses the standard crossover method with three objectives. We implemented Avg Ensemble and MMOEA clustering methods and the code for MOCK was provided by the authors.

Table I, II and III show the percentage values of CA, F1 and RI respectively computed on 10 different two view datasets. The v1 means view one and "v1+v2" means that view one and view two were concatenated in a single feature matrix. The 2view means the two views v1 and v2 were separately used and two feature matrices were constructed in the clustering process. The bold values indicate the highest value. All three tables show a general trend that our new method, MDC, outperforms other clustering methods in terms of CA, F1 and RI on all views.

Table IV provides the percentage improvements of MDC

[1]http://personalpages.manchester.ac.uk/mbs/julia.handl/mock.html

TABLE IV: Improvement of clustering quality of MDC compare to MMOEA

Dataset	CA Improvement			F1 Improvement			RI Improvement		
	v1	v1+v2	2view	v1	v1+v2	2view	v1	v1+v2	2view
D1	0.000	0.000	1.525	0.000	0.000	0.000	0.647	0.960	0.293
D2	0.000	0.000	1.139	0.000	0.000	0.000	0.437	0.255	0.401
D3	2.520	1.955	6.238	0.353	0.403	1.823	0.994	3.994	9.834
D4	**3.075**	3.668	4.902	0.776	2.235	3.192	0.694	**5.312**	**16.711**
D5	2.936	**3.357**	7.273	1.447	1.806	6.609	**3.291**	1.132	8.599
D6	1.507	2.645	3.552	**1.518**	**4.415**	5.771	3.166	4.376	8.346
D7	0.194	0.041	5.069	1.397	3.678	4.449	1.750	0.468	5.708
D8	0.000	0.568	5.006	1.076	1.696	2.492	2.675	2.823	3.678
Citeseer	2.220	3.635	**10.637**	3.924	3.348	**11.530**	2.623	2.891	10.333
Cora	0.707	2.347	9.145	1.543	1.668	2.138	1.675	1.384	9.679

TABLE V: Statistical significance test based on the values of Clustering Accuracy, F1-measure and Rand Index computed on all datasets

	Avg-Ensemble		MOCK		MMOEA		
	v1	v1+v2	v1	v1+v2	v1	v1+v2	2view
p-value on CA	0.0015	0.0008	0.0025	0.0009	0.0154	0.0057	0.0001
p-value on F1	0.0017	0.0185	0.0056	0.0447	0.0042	0.0026	0.0038
p-value on RI	0.0003	0.003	0.0004	0.008	0.0005	0.0021	0.0011

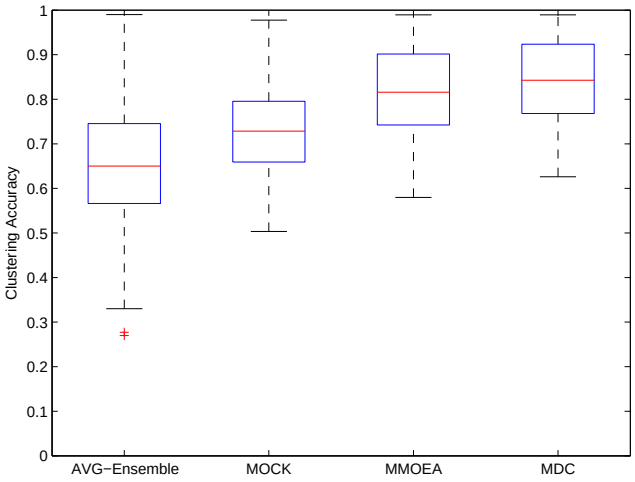

Fig. 6: CA on combined dataset

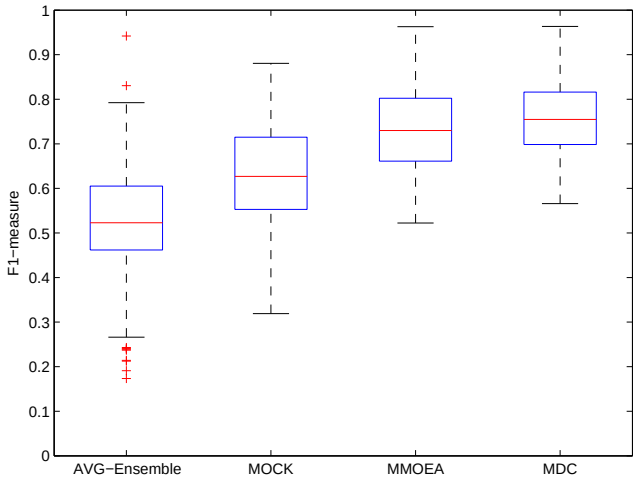

Fig. 7: F1 on combined dataset

over MMOEA. The percentage values are calculated in terms of CA, F1 and RI on 10 different datasets. The bold values represent the highest improvement achieved by MDC. In general, MDC performs equally well for a few cases, but most of the time shows a reasonable improvement over MMOEA.

D. Comparison on Three View Datasets

The results of three view/combined dataset of AMBIENT, MORESQUE and ODP-239 are depicted in Figure 6, 7 and 8 in terms of the boxplot of CA, F1 and RI values. The three views were concatenated for Avg-Ensemble and MOCK and used separately for MMOEA and MDC. Our method MDC has better mean values as compared to Avg-Ensemble, MOCK and MMOEA in terms of CA, F1 and RI values. Similar results were also observed on other views.

V. Discussion

A. Statistical Analysis

We performed the pairwise Wilcoxon statistical significance test [26] on all datasets for CA, F1 and RI values. We used CA, F1 and RI values of MDC as a control group and was compared them individually with the values of Avg-Ensemble, Mock and MMOEA. Table V shows the p-values of the statistical test performed on CA, F1 and RI values of 10 different datasets(D1-D8, Citeseer and Cora). The p-values of statistical test performed on CA, F1 and RI values

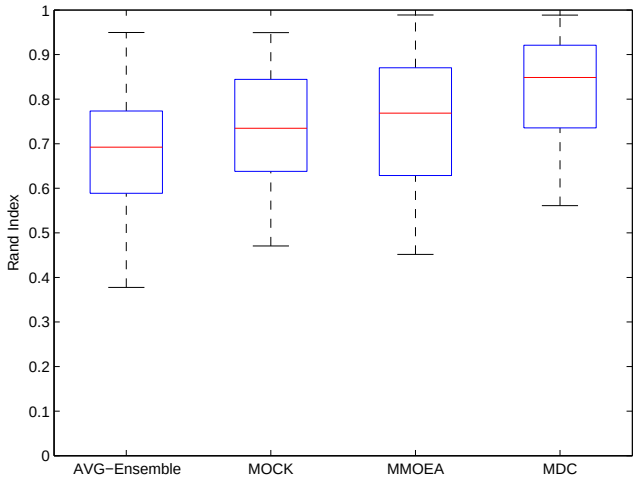

Fig. 8: RI on combined dataset

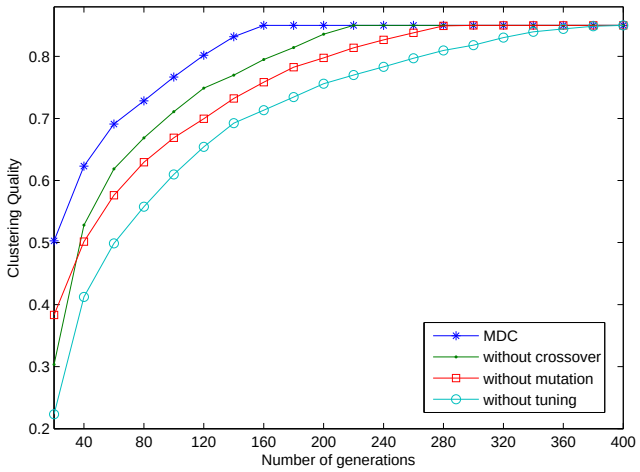

Fig. 9: Analysis of MDC

of combined dataset (having 397 queries) are 0.0016, 0.0002 and 0.0001 respectively. We used α = 0.05 for all statistical test and the results showed that our method MDC has a statistically significant improvement as compared to other clustering methods.

B. General Analysis

We also implemented single objective methods GA-1 and GA-2 based on our approach MDC. The GA-1 uses Obj_1 and GA-2 uses Obj_2 only. Table VI shows F1 score computed on 10 different datasets for GA-1, GA-2 and MDC for comparison. MDC outperformed single objective clustering methods GA-1, GA-2. This analysis was performed to see if generating diverse clustering is the only reason for the better performance and we found that using multi-objective approach we can improve the results.

We further performed component analysis of MDC by removing different evolutionary steps and analyzing the results. Figure 9 shows the clustering quality (Average CA) on y-axis and number of generations on x-axis. The results were computed by removing crossover, mutation and tuning steps one by one from MDC method and producing the clustering solution. With all features, MDC converged around 160 generations, however when the tuning step was removed the method was converged around 375 generations. The tuning step played a vital role in improving the convergence speed.

VI. Conclusion

This paper presented a new multi-objective clustering ensemble method (MDC) based on SPEA-II for clustering documents. MDC uses the concept of multiple views to generate multiple clustering solutions with diversity. Then using evolutionary process, it generates a better clustering solution by select a diverse set of clustering solution and then selecting high-quality clusters to derive a final clustering solution.

MDC outperformed other recent clustering methods (state-of-the-art non evolutionary and evolutionary clustering methods) and its single objective variants. In future, we would like to investigate the automatic detection of the multiple views in documents and the scalability of MDC.

References

[1] S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 03, pp. 337–372, 2011.

[2] J.-P. Barthélemy and B. Leclerc, "The median procedure for partitions," *DIMACS series in discrete mathematics and theoretical computer science*, vol. 19, pp. 3–34, 1995.

[3] M. H. Law, A. P. Topchy, and A. K. Jain, "Multiobjective data clustering," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–424.

[4] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[5] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 1, pp. 56–76, 2007.

[6] E. E. Korkmaz, J. Du, R. Alhajj, and K. Barker, "Combining advantages of new chromosome representation scheme and multi-objective genetic algorithms for better clustering," *Intelligent Data Analysis*, vol. 10, no. 2, pp. 163–182, 2006.

[7] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 5, pp. 1506–1511, 2007.

[8] L. I. Kuncheva and J. C. Bezdek, "Selection of cluster prototypes from data by a genetic algorithm," in *Proc. 5th European Congress on Intelligent Techniques and Soft Computing, Aachen, Alemanha*, 1997, pp. 1683–1688.

[9] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, no. 8, pp. 841–847, 1991.

[10] J. Handl and J. Knowles, "Evidence accumulation in multiobjective data clustering," in *Evolutionary Multi-Criterion Optimization*. Springer, 2013, pp. 543–557.

[11] D. Dutta, P. Dutta, and J. Sil, "Simultaneous feature selection and clustering with mixed features by multi objective genetic algorithm," *International Journal of Hybrid Intelligent Systems*, vol. 11, no. 1, pp. 41–54, 2014.

[12] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes," *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 5, pp. 991–1005, 2009.

[13] S. J. Nanda and G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," *Swarm and Evolutionary Computation*, 2014.

TABLE VI: Comparison of F1 score of GA-1, GA-2, MDC (SPEA-II) and MDC (NSGA-II) computed on 10 different multi-view datasets

Dataset	GA-1		GA-2		SPEA-II (MDC)		
	v1	v1+v2	v1	v1+v2	v1	v1+v2	2view
D1	71.94	72.94	71.01	72.94	**91.95**	**91.68**	**92.01**
D2	76.14	76.14	76.99	76.34	**92.61**	**92.69**	**93.28**
D3	57.01	58.41	68.19	69.12	**91.03**	**92.11**	**93.82**
D4	59.38	61.55	69.11	70.33	**86.99**	**89.18**	**92.78**
D5	61.02	61.96	63.12	63.76	**70.12**	**71.61**	**75.82**
D6	66.73	48.13	67.53	50.11	**68.21**	**70.72**	**74.05**
D7	65.52	67.96	72.12	73.01	**63.87**	**69.63**	**74.42**
D8	75.58	76.83	75.88	77.63	**75.12**	**76.74**	**78.13**
Citeseer	60.43	73.06	72.23	74.34	**61.18**	**61.73**	**68.29**
Cora	70.21	73.92	76.12	74.23	**60.54**	**62.17**	**63.54**

[14] S. Bickel and T. Scheffer, "Multi-view clustering." in *ICDM*, vol. 4, 2004, pp. 19–26.

[15] A. Wahid, X. Gao, and P. Andreae, "Exploiting user queries for search result clustering," in *Web Information Systems Engineering–WISE 2013*. Springer, 2013, pp. 111–120.

[16] ——, "Multi-view clustering of web documents using multi-objective genetic algorithm," in *Evolutionary Computation (CEC), 2014 IEEE Congress on*. IEEE, 2014, pp. 2625–2632.

[17] ——, "Multi-objective multi-view clustering ensemble based on evolutionary approach," in *To appear in Evolutionary Computation, 2015. CEC'15. IEEE Congress on*. IEEE, 2015.

[18] E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: Improving the strength pareto evolutionary algorithm," 2001.

[19] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. P. L. F. De Carvalho, "A survey of evolutionary algorithms for clustering," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 2, pp. 133–155, 2009.

[20] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining Text Data*. Springer, 2012, pp. 77–128.

[21] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artificial Intelligence*, 2012.

[22] N. Nguyen and R. Caruana, "Consensus clusterings," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 607–612.

[23] D. Crabtree, X. Gao, and P. Andreae, "Improving web clustering by cluster selection," in *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*. IEEE, 2005, pp. 172–178.

[24] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[25] N. Iam-on and S. Garrett, "Linkclue: A matlab package for link-based cluster ensembles," *Journal of Statistical Software*, vol. 36, no. i09, 2010.

[26] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.