Knowl Inf Syst manuscript No.
(will be inserted by the editor)

rFILTA: Relevant and Non-Redundant View Discovery from Collections of Clusterings via Filtering and Ranking

Yang Lei[1] · **Nguyen Xuan Vinh**[1] · **Jeffrey Chan**[2] · **James Bailey**[1]

Received: Jun 26, 2015 / Revised: Jul 16, 2016 / Accepted: Oct 22, 2016

Abstract Meta-clustering is a popular approach for finding multiple clusterings in the dataset, taking a large number of base clusterings as input for further user navigation and refinement. However, the effectiveness of meta-clustering is highly dependent on the distribution of the base clusterings and open challenges exist with regard to its stability and noise tolerance. In addition, the clustering views returned may not all be relevant, hence there is open challenge on how to rank those clustering views. In this paper we propose a simple and effective filtering algorithm that can be flexibly used in conjunction with any meta-clustering method. In addition, we propose an unsupervised method to rank the returned clustering views. We evaluate the framework (rFILTA) on both synthetic and real world datasets, and see how its use can enhance the clustering view discovery for complex scenarios.

Keywords Clustering · Meta-Clustering · Multiple Clusterings · Clustering Visualization · Clustering Filtering · Clustering Ranking

1 Introduction

Clustering is one of the most important unsupervised techniques for discovering unknown patterns and grouping in the dataset. Many clustering methods focus on obtaining one single 'best' solution by optimizing a pre-defined criterion [21, 29, 28]. There are two limitations with this: firstly, data can be multi-faceted in

[1]Yang Lei
E-mail: yalei@student.unimelb.edu.au
[1]Nguyen Xuan Vinh
E-mail: vinh.nguyen@unimelb.edu.au
[2]Jeffrey Chan
E-mail: jeffrey.chan@rmit.edu.au
[1]James Bailey
E-mail: baileyj@unimelb.edu.au
[1]Department of Computing and Information Systems, University of Melbourne, Australia
[2]School of Science (Computer Science), RMIT University, Australia
Jeffrey Chan conducted part of this work while at the University of Melbourne.

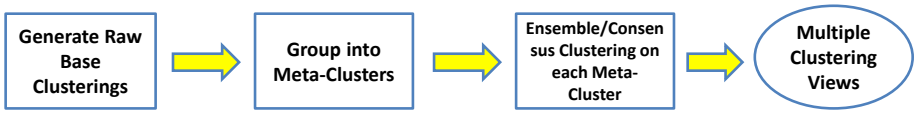
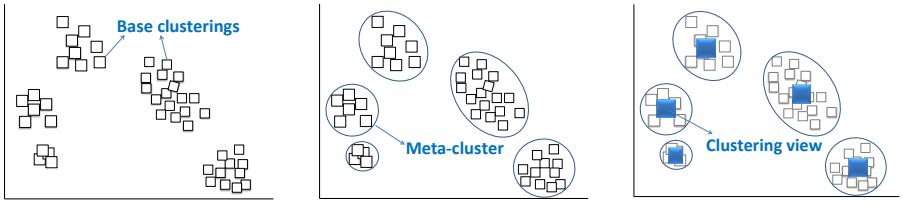


Fig. 1: The meta-clustering framework.



(a) The base clusterings. Each symbol represents a base clustering.

(b) The base clusterings are grouped into meta-clusters.

(c) One clustering view is generated for each meta-cluster.

Fig. 2: Illustrative example for meta-clustering process.

nature. Particularly when the datasets are large and complex, there may be several useful clusterings that exist, not just one. Secondly, users may be seeking different perspectives on the same dataset, requiring multiple clustering solutions. This has stimulated considerable recent research on the topic of *multiple clustering analysis* [4].

Multiple clustering analysis aims to discover a set of reasonable and distinctive clustering solutions from the same dataset. Many methods have been proposed on this topic and one very popular technique is meta-clustering [5],[43]. Meta-clustering generates a large number of base clusterings using different approaches [5]: running different clustering algorithms, running a specific algorithm several times with different initializations, or using random feature weighting in the distance function. These base clusterings may then be meta-clustered into groups (Figure 2b). Then, (base) clusterings within the same group can be combined using consensus (ensemble) clustering to generate a consensus view of that group (Figure 2c). This results in one or more distinctive clustering views of the dataset, each offering a different perspective or explanation. The general procedure of the meta-clustering is described in Figure 1.

1.1 Motivation for Filtering

A major drawback and challenge with meta-clustering is that its effectiveness is highly dependent on the quality and diversity of the generated base clusterings. Specifically, if the base clusterings are of low quality, then the ensemble step will be influenced by such clusterings and may produce low quality clustering views. In addition, if there are redundant, noisy base clusterings that are similar to one or more of the clustering views, then it is possible that some of the distinct views are

(a) 4 representative clustering views generated from the unfiltered base clusterings. The similarity scores between each clustering view with the two ground truth views respectively are presented at the top of each clustering view.

(b) 2 clustering views generated from the filtered base clusterings. The similarity scores between each clustering view with the two ground truth views respectively are presented at the top of each clustering view.

Fig. 3: Clustering views generated from the unfiltered and filtered base clusterings on CMUFace dataset.

mistakenly merged into one, resulting in the loss of interesting clustering views. This can occur if the base clusterings representing two distinct views are connected via a chain of noisy but similar base clusterings. The grouping algorithm may then mistakenly group all of these base clusterings into one meta-cluster (Figure 2b) and subsequently one clustering view will be produced by the ensemble step (Figure 2c) when it finds a consensus view from the merged meta-cluster. In this way, users may miss some interesting clustering views.

We have experienced these problems in our experiments of both synthetic and real world datasets. To illustrate, we use an example from a real dataset (shown in Figure 3). The CMUFace dataset, which contains images of three different persons along with different poses (left, front and right), consists of two reasonable clustering views, Person and Pose[1]. From the CMUFace dataset, we generate a set of (raw) base clusterings (with $k = 3$ clusters) using a number of standard base clustering generation algorithms (see Section 9.1). Some of the generated base clusterings contain the Person or Pose clustering views, so it should be possible to recover/discover both views. We then applied meta-clustering on the generated (raw) base clusterings and found 23 clustering views. Due to the limitation of space, we show four representative clustering views in Figure 3a. In Figure 3a, each row is a clustering view of three clusters, where each cluster is shown as the

[1] Please refer to Section 9.5.1 for more details about the dataset and experiments.

mean of all the images in it. Above each clustering view, we show the similarity score[2], ranging within $[0, 1]$, between this clustering view and the two ground truth clustering views, i.e., Person and Pose views, respectively. The larger the value is, the more similar the clustering view is to that ground truth clustering view.

As we can see from Figure 3a, only one of the ground truth clustering views is discovered, i.e., Person view (the first clustering view), with a similarity score of 0.83. However, we could not discover the Pose view from the other clustering views. In addition, many of the clustering views are of poor quality and do not correspond to any underlying view, e.g., the other three shown clustering views in Figure 3a. The reason for this is that some of the generated base clusterings are noisy and of low quality (i.e., not of either ground truth clustering views) and/or form bridges between the base clusterings of the two clustering views. This large amount of noise causes many redundant and poor quality clustering views found, and also the pose view being lost in the noise (e.g., the third clustering view/row in Figure 3a has a significant number of pose images in their clusters, but the bridging base clusterings have caused some person view images to be merged with it.) These observations stimulate the following question, which is the motivation for the proposed filtering method - *Can we filter out the redundant/similar and noisy base clusterings to avoid discovering poor quality views or missing out on significant ones?* Figure 3b provides an example of the benefits of filtering. It shows the two clustering views, i.e., Person and Pose, generated using a filtered set of base clusterings as input (we will explain our filtering approach in Section 4). More specifically, after filtering out the poor quality base clusterings, we avoid clustering views of poor quality. In addition, filtering the redundant base clusterings helps to expose the other reasonable clustering view, i.e., pose view (Figure 3b). More examples will be presented in the experiments (Section 9).

1.2 Motivation for Ranking

Another challenge about meta-clustering is the large number of generated clustering views. Depending on the datasets and the base clusterings generation mechanism, we may produce a large number of clustering views. It will be time consuming to examine them all. Our filtering step can help reduce the number of clustering views by removing out the ones of poor quality. However, depending on the complexity of the datasets, the generated base clusterings and the different requirements of users, we do not know how many potentially interesting clustering views exist. There may be still a large number of potentially interesting clustering views generated after filtering. Hence, it will be helpful to rank these clustering views based on importance and diversity, and provide users with the top K ones, to facilitate their analysis job.

A question that may arise, 'can we obtain high quality and diverse clustering views with ranking alone, i.e., without filtering?' The answer is no. The ranking step helps solve the problem about the large number of clustering views. However, for the problems introduced previously (refer to Section 1.1), e.g., missing interesting clustering views, cannot be solved by ranking. Thus, we need both filtering and ranking to help us get the good quality and diverse clustering views.

[2] The similarity between two clusterings are measured by *adjusted mutual information* (AMI), which will be introduced in Section 4.1.

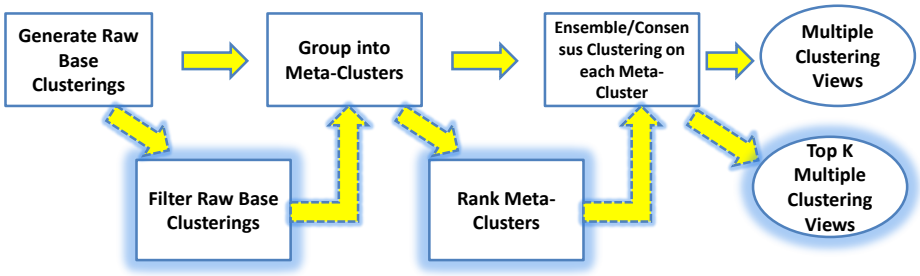


Fig. 4: The meta-clustering framework with our additional, proposed filtering and ranking steps highlighted with shaded square.

1.3 Our Contribution

To solve all the problems described above, we propose our new approach, ranked and filtered meta-clustering (rFILTA), aiming at detecting multiple high quality and distinctive clustering views by filtering, ranking and analyzing a given set of base clusterings.

Algorithmically, we propose an information theoretic criterion to perform the filtering. In addition, we show how to employ a visual method to automatically determine the meta-clusters within the filtered base clusterings. Then, we can rank these meta-clusters in terms of their importance measured by the proposed heuristic criteria. Finally, we perform consensus clustering on the returned top K meta-clusters to produce the top K clustering views. Figure 4 shows the whole process. The novelty of our approach lies in the addition of a filtering step and a ranking step to the existing meta-clustering framework [5,43], which are highlighted as the shaded square boxes in Figure 4. Our focus is on investigating how to filter the given raw base clusterings to generate a set of better clustering views, in terms of quality and diversity, compared to the unfiltered meta-clustering and also on ranking the clustering views in terms of their importance. We assume that we are given a set of base clusterings. The generation of appropriate base clusterings, which has been considered by numerous existing work in the literature [21, 5,33,43,13,1,17], is outside the scope of this paper. This is not a large limitation as any existing meta-clustering generation techniques can be used. An important advantage of our method is that the filtering step and the ranking step are independent of the other steps in this framework and thus may be easily integrated with them.

A preliminary version of this work first appeared in [24]. This work differs from previous work in that:

- We identify the desirability of a ranking mechanism to assist in selecting a small number of informative views.
- We propose several heuristic ranking schemes for ranking the multiple clustering views in terms of their importance, to further assist users in their analysis with a large number of clustering views.

- We modify the parametrization of the objective function in the filtering step, in which the tradeoff parameter β now ranges within $[0, 1]$ instead of $[0, \infty]$, making it easier for users to tune.
- We evaluate and demonstrate the performance of our new rFILTA framework on 8 datasets (in addition to the 1 synthetic and 2 real world datasets in the preliminary work, we add 1 synthetic and 4 real world datasets).

The rest of the paper is organized as follows. We review the related work in Section 2. Then, the general outline of our rFITLA framework is shown in Section 3. Next we present the filtering approach detailedly in Section 4. The grouping method used in our rFILTA framework is introduced in Section 5. We describe the multiple heuristic ranking schemes in Section 6. Then, we introduce the ensemble method used in this work in Section 7. The time complexities of different steps involved in rFILTA framework are introduced in Section 8. Finally, we present the exhaustive experimental results and analysis on 8 datasets in Section 9 and conclude in Section 10.

2 Related Work

Our research is related to several topics: meta-clustering, alternative clustering and cluster ensemble or consensus clustering.

Meta-Clustering aims to find multiple clustering views by generating and evaluating a large set of base clusterings. In work [5], it first generates these base clusterings by either random initialization or random feature weighting. Then, it groups these base clusterings into multiple meta-clusters and then presents these meta-clusters to the users for evaluation. Based on this idea, Zhang and Li [43] proposed a method that extend [5] with consensus clustering in order to capture multiple views. Work in [33] proposed a sampling method for discovering a large set of good quality base clusterings. After that, the k-center [15] clustering method is used to select the k most dissimilar solutions as the views. In contrast to rFILTA, the existing meta-clustering methods are highly dependent on the quality and diversity of the base clusterings for generating multiple high quality and diverse views.

Alternative Clustering discovers high quality and dissimilar views via searching in the clustering space guided by criteria about what constitutes an alternative. One may discover alternatives either iteratively or simultaneously [3, 6, 11, 22, 39, 20, 31, 9, 10]. See [4] for a review. Compared with meta-clustering, alternative clustering is more efficient for discovering alternative views. However, it restricts the definition of an alternative to certain objective functions, which may cause the search process to miss some interesting clustering views, due to mismatches between the objective function and the underlying view structure. It can be difficult to define an objective function characterizing what is an alternative, especially in the initial period of data exploratory analysis when there is little information about the data available.

Cluster Ensemble or Consensus Clustering combines a collection of partitions of data into a single solution which aims to improve the quality and stability of individual clusterings [37, 38, 27, 14, 41, 16]. However, instead of combining all the available clusterings into one single solution, it has been demonstrated that

a better clustering can often be achieved by combining only a part of the available solutions [13,1,26], that is the **cluster ensemble selection problem**. It has been shown that quality and diversity are two important factors which will influence the performance of cluster ensemble [13,17,26]. Cluster ensemble and the cluster ensemble selection methods typically focus on discovering a single high quality solution from a collection of clusterings, rather than multiple solutions.

Our proposed framework in Figure 4 combines all of the above clustering paradigms. The critical difference between our work compared to the others is that we place each clustering paradigm into its most relevant place. In particular, we employ alternative clustering as one of the mechanisms for generating diverse base clusterings. Alternative clustering employs objective functions to guide the search process, thus it may discover alternative clustering views faster compared to meta-clustering which employs a random clustering generation scheme (such as random initialization or random feature weighting). However, if the objective function defined in the alternative clustering cannot characterize the underlying structure of the dataset appropriately, it cannot discover the alternative clustering view [3]. On the other hand, meta-clustering can cover the space of clusterings more comprehensively compared to alternative clustering, by flexibly employing different means of generation. Thus, we take alternative clustering method as one of the generation methods. Finally, we propose to group the clusterings and generate the consensus view for each group via consensus clustering. This is a more flexible approach than generating a single consensus view for the whole set of base clusterings, as the base clusterings may reflect very different structures of the data and thus may not be reasonably combined to produce a single consensus view. In summary, our rFILTA framework incorporates the strengths of the related techniques to improve upon the existing meta-clustering methods.

3 Notations and rFILTA Framework

Let us firstly introduce the notations used in this paper. Let $X = \{x_1, \ldots, x_n\}$ be a set of n objects, where $x_i \in \mathbb{R}^{\mathbf{d}}$. These objects can be grouped into clusters (sets of objects). A clustering C is a hard partition of X, denoted by $C = \{c_1, \ldots, c_k\}$, where c_i is a cluster and $c_i \cap c_j = \emptyset, \bigcup c_i = X$. We denote the space of possible clusterings on X as $\mathcal{P}_{\mathcal{X}}$. We use $\mathcal{C}$ to denote a set of (base) clusterings, i.e., $\mathcal{C} = \{C_1, \ldots, C_l\}$. Let a set of clustering views be denoted by $\mathcal{V} = \{V_1, \ldots, V_R\}$, where a clustering view V_i is a clustering on X, $V_i \in \mathcal{P}_{\mathcal{X}}$. Even though a clustering view is just a clustering, we use the view nomenclature to distinguish between the initial base clusterings and the final, returned clusterings (the set of clustering views) at the end of the meta-clustering process.

The rFILTA framework consists of a number of steps, illustrated in Figure 4 and Algorithm 1. In the following sections, we will describe each of these steps.

4 Filtering Base Clusterings

The quality of a clustering C is measured by a function $Q(C)$: $\mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}^+$, and the diversity between two clusterings can be computed according to a similarity

[3] We demonstrate this point in the experiments part, refer to Figure 21

Algorithm 1 Framework of rFILTA

Input:

Generated base clusterings $\mathcal{C} = \{C_1, \ldots, C_l\}$
K, the number of returned clustering views
L, the number of selected base clusterings during filtering step
β, $\beta \in [0, 1]$, the tradeoff parameter which balance the quality and diversity during filtering

Output:

The top K clustering views, $\mathcal{V}'_K = \{V'_1, \ldots, V'_K\}$

1: $\mathcal{C}' \leftarrow Filtering(\mathcal{C}, L, \beta)$, where $\mathcal{C}' = \{C'_1, \ldots, C'_L\}$ (Section 4)
2: $\mathcal{C}_{mc} \leftarrow Grouping(\mathcal{C}')$, where $\mathcal{C}_{mc} = \{\mathcal{C}_{mc_1}, \ldots, \mathcal{C}_{mc_R}\}$, $\mathcal{C}_{mc_i} \cap \mathcal{C}_{mc_j} = \emptyset, \bigcup \mathcal{C}_{mc_i} = \mathcal{C}'$ (Section 5)
3: $\mathcal{C}'_{mc} \leftarrow Ranking(\mathcal{C}_{mc}, K)$, where $\mathcal{C}'_{mc} = \{\mathcal{C}'_{mc_1}, \ldots, \mathcal{C}'_{mc_K}\}, \mathcal{C}'_{mc} \subset \mathcal{C}_{mc}$ (Section 6)
4: $\mathcal{V}'_K \leftarrow Consensus(\mathcal{C}'_{mc})$, where $\mathcal{V}'_K = \{V'_1, \ldots, V'_K\}$ (Section 7)
5: **return** $\mathcal{V}'_K$

measure $Sim(C_i, C_j)$: $\mathcal{P}_\mathcal{X} \times \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}^+$. The filtering problem can be formalized as follows.

Problem Definition *Given a set of raw base clusterings $\mathcal{C} = \{C_1, \ldots, C_l\}$, we seek a set of clustering views $\mathcal{V} = \{V_1, \ldots, V_R\}$ generated from $\mathcal{C}$, such that, $\sum_{V_i \in \mathcal{V}} Q(V_i)$ is maximized and $\sum_{V_i, V_j \in \mathcal{V}, i \neq j} Sim(V_i, V_j)$ is simultaneously minimized.*

We solve this problem by selecting a subset of clusterings $\mathcal{C}'$, which are of high quality and diversity, from the given raw base clusterings $\mathcal{C}$. The quality and diversity of base clusterings have a big impact on the quality and diversity of the extracted clustering views at last. Next we discuss the quality and diversity criteria for clusterings.

4.1 Clustering Quality and Diversity Measures

We employ an information theoretic criterion, namely the mutual information for measuring both clustering quality and diversity. As a clustering quality measure, mutual information is a well-known criterion for clustering discovery, which can discover both linear and non-linear clusterings [8]. For measuring similarity between clusterings, mutual information can detect linear or non-linear relationship between random variables [39]. More specifically, the quality of a clustering C is measured by the amount of shared information with the data X, i.e., $I(X; C)$. The more information that is shared, the better that a clustering models the data. In contrast, the mutual information between two clusterings $I(C_i; C_j)$ quantifies their similarity. The less mutual information shared between the clusterings, the more dissimilar they are. There are many mutual information variations (see [39] for a survey). We choose to utilize the *Adjusted Mutual Information* (AMI) [40], which is an adjusted-for-chance version of the normalized mutual information[4] [37]

[4] The normalized version of mutual information for scaling mutual information to $[0, 1]$.

for measuring the similarity between two clusterings. We selected AMI as its value lies between the interpretable range of 0 to 1 and it uses a principled approach to normalize to such a range.

The average quality of the selected set of base clusterings can be optimized as:

$$\max_{\mathcal{C}'} \left\{ \frac{1}{|\mathcal{C}'|} \sum_{C_i \in \mathcal{C}'} I(X; C_i) \right\} \equiv \min_{\mathcal{C}'} \left\{ \frac{1}{|\mathcal{C}'|} \sum_{C_i \in \mathcal{C}'} H(X|C_i) \right\} \tag{1}$$

where the right hand side results from $I(X;C) = H(X) - H(X|C)$ and $H(X)$ is a constant (where H$(\cdot)$ is the Shannon entropy function). Computation of the mutual information $I(X;C)$ requires the joint density function, $p(X,C)$, which is difficult to estimate for high dimensional data. Instead of directly estimating the joint densities, we may use the meanNN differential entropy estimator for computing the conditional entropy $H(X|C)$ [12], due to its desirable properties of efficiently estimating density functions in high dimensional data and being parameterless. It is defined as

$$H(X|C) \approx \sum_{j=1}^{n_c} \frac{1}{n_j - 1} \sum_{i \neq l | c_i = c_l = j} \log ||x_i - x_j|| \tag{2}$$

The diversity can be optimized by minimizing the average similarity between clusterings, as:

$$\min_{\mathcal{C}'} \left\{ \frac{1}{|\mathcal{C}'|^2} \sum_{C_i, C_j \in \mathcal{C}'} AMI(C_i; C_j) \right\}$$

The AMI between two clusterings C_i and C_j is defined as:

$$AMI(C_i; C_j) = \frac{I(C_i; C_j) - E\{I(C_i; C_j)\}}{\max\{H(C_i), H(C_j)\} - E\{I(C_i; C_j)\}} \tag{3}$$

where the $E\{\cdot\}$ is the expected value of mutual information $I(C_i; C_j)$, and $H(C_i) = -\sum_{u \in C_i} p(u) \log p(u)$, is the entropy of the clustering C_i. The AMI is 1 when the two clusterings are identical and 0 when any commonality between the clusterings is due to chance. The mutual information between two clusterings C_i and C_j is computed directly from their contingency table:

$$I(C_i; C_j) = \sum_{u \in C_i} \sum_{v \in C_j} p(u, v) \log \frac{p(u, v)}{p(u)p(v)} \tag{4}$$

where $p(v)$ is the fraction of data points in cluster v, and $p(u, v)$ is the fraction of points belonging to cluster u in C_i and v in C_j.

4.2 Filtering Criterion and Incremental Selection Strategy

We wish to select a subset of base clusterings, $\mathcal{C}'$, to achieve high quality and diversity simultaneously. Inspired by the mutual information based feature selection literature [32] which maximizes feature relevancy while minimizing feature redundancy, we propose a clustering selection criterion which combines the quality and diversity of clusterings:

$$\min_{\mathcal{C}' \subset \mathcal{C}, |\mathcal{C}'|=L} \left\{ \frac{\beta\beta_0}{|\mathcal{C}'|} \sum_{C_i \in \mathcal{C}'} H(X|C_i) + \frac{1-\beta}{|\mathcal{C}'|^2} \sum_{C_i, C_j \in \mathcal{C}', i \neq j} AMI(C_i; C_j) \right\} \quad (5)$$

where L is a user defined parameter specifying the number of base clusterings $\mathcal{C}'$ to be selected, and $\beta \in [0, 1]$ is a trade-off parameter that balances the emphasis put on the quality and diversity during selection. When $\beta = 0.5$, we put same emphasis on the quality and diversity. The influence of β is discussed in the experiments Section 9.3.2. To make sure the first term is on the same scale as the first term, we rescale the $H(X|C_i)$ to $[0, 1]$ by multiplying it with $\beta_0 = \frac{H(X|C_i) - min\{H(X|C_i)\}}{max\{H(X|C_i)\} - min\{H(X|C_i)\}}$. Thus, our selection method aims to select L base clusterings $\mathcal{C}'$ from the given raw base clusterings $\mathcal{C}$, optimizing the dual-objective criterion in Equation (5).

A simple incremental search strategy can be used to select a good subset $\mathcal{C}'$ for the criterion (5) as follows. Initially, we select the clustering solution with the highest quality among the given clusterings $\mathcal{C}$. Then, we incrementally select the next solution from the set $\mathcal{C} \setminus \mathcal{C}'$ as:

$$\arg\min_{C_i \in \mathcal{C} \setminus \mathcal{C}'} \left\{ \beta\beta_0 H(X|C_i) + \frac{1-\beta}{|\mathcal{C}'|} \sum_{C_j \in \mathcal{C}'} AMI(C_i; C_j) \right\} \quad (6)$$

with the aim of selecting the next clustering with high quality and small average similarity with the selected ones in $\mathcal{C}'$. This process repeats until we reach the L desired number of base clusterings.

5 Discovering the Meta-Clusters

We have obtained a filtered set of base clusterings after performing the filtering process. Next we group them into clusters at the meta level and then perform ensemble clustering on each meta-cluster for view generation. We first explain the measure used to compute the similarity between the base clusterings, then explain a visualization technique called iVAT for determining the potential number of meta-clusters. We then introduce a method that combines with iVAT to automatically determine the appropriate number of meta-clusters and performs the grouping, and finally describe how to obtain the views from the meta-clusters.

Measuring the Similarity between Clusterings: In order to divide the selected clusterings into groups, we need a similarity measure to compare clusterings. Several measures of clustering similarity have been proposed in the literature [21]. Here we use the AMI for measuring the similarity between clusterings. The distance between two clusterings is then $1 - AMI(C_i; C_j)$.

Grouping the Base Clusterings into Meta-Clusters: After filtering the base clusterings to obtain $\mathcal{C}'$, we compute the pairwise dissimilarity matrix between all members of $\mathcal{C}'$ as a prelude to grouping them into meta-clusters. There are two

challenges for this grouping step: a) determining the number of relevant meta-clusters; and b) partitioning the clusterings into meta-clusters. Next, we will describe a visualization technique for assessing the number of meta-clusters in a set of base clusterings. Then, an automatic method for determining the number of meta-clusters and partitioning the clusterings into meta-clusters will be presented.

The VAT method [42] is a visualization tool for cluster tendency assessment. By reordering a pairwise dissimilarity matrix of a set of data objects, it can reveal the hidden clustering structure of the data by visualizing the intensity image of the reordered dissimilarity matrix. The number of clusters in a set of data objects can be visually identified by the number of "dark blocks" displayed along the diagonal of the VAT image. In this work, we use the iVAT [42,18] method which is an advanced version of VAT method, in terms of presenting clearer blocks in the images of the reordered dissimilarity matrix. Each clustering can be taken as a data object, and we utilize the iVAT method to visualize the number of potential meta-clusters.

For grouping the set of clusterings, existing research uses hierarchical clustering [5],[43]. However, the hierarchical clustering approach does not automatically determine the number of clusters. Hence, we propose an alternative, CLODD [19], the clustering method which automatically extracts the number of clusters and produces a hard partition of the data objects. We choose CLODD, as it also works on reordered dissimilarity matrices generated by the iVAT method. We believe these two methods are well complementary. Nevertheless, we stress that the rFILTA framework is not restricted to any particular grouping method and users can choose the one they prefer according to their requirements and knowledge. As mentioned above, there will be dense blocks along the diagonal of this ordered dissimilarity matrix if clusters exist in this set of clusterings. The CLODD algorithm discovers the number of meta-clusters and produces a hard partition of these clusterings by optimizing an objective function which assesses the dense diagonal block structures of the reordered dissimilarity matrix. At the end of this step, we have multiple meta-clusters generated.

For the CLODD method, there are two involved parameters to be set, namely NC_{min} and NC_{max}, which are the minimum and maximum number of meta-clusters. We set $NC_{min} = 1$ and $NC_{max} =$ the number of base clusterings to be grouped. For all other parameters involved in the CLODD method, we use the values suggested in the original work [19] as default.

6 Meta-Cluster Ranking

When the generated base clusterings are widely distributed in the clustering space and the tradeoff parameter β is chosen small, many clustering views may be produced, even after filtering. When there are a lot of clustering views, it will be helpful if we can rank and show the top K clustering views for users to analyze. Examining many irrelevant clustering views, which is a possibility when there is no ranking, is time consuming and frustrating. The challenge is how to rank the clustering views. There are different definitions characterizing what is a good clustering view, according to different requirements of different users. It is hard to define a criterion to satisfy all these different requirements. Moreover, there is no standard 'right' ranking for us to learn from.

In this section, we propose several heuristic ranking schemes for ranking the meta-clusters based on their characteristics. Then we will apply ensemble clustering on the returned top K meta-clusters to produce the top K clustering views. These schemes are reasonable options in terms of different considerations, and users can choose from them according to their requirements.

Recall that a meta-cluster is a set of clusterings which ideally correspond to a clustering view. A clustering view is a clustering that represent a meta-cluster. Different from traditional cluster evaluation, we have to consider the quality of the meta-cluster as they consist of clusterings as members. The quality of the members of a meta-cluster has a big impact on the goodness of its corresponding clustering view. Next we present several properties that can be considered for measuring the goodness of a meta-cluster.

1. Cohesion and Separation
   Similar to cluster evaluation, we can take the terms of cohesion and separation, which are used for measuring the goodness of a cluster, for measuring the goodness of a meta-cluster. The more compact a meta-cluster, the more similar the clusterings within this meta-cluster. This indicates the clusterings within this meta-cluster can be repeatedly found by some of the available clustering generation methods. Thus, this meta-cluster is more likely to correspond to a reasonable clustering view. If the meta-cluster is separated well from other meta-clusters, it indicates this meta-cluster is different from the others and corresponds to a distinctive clustering view. Based on these ideas, we build an *Meta-Cluster Cohesion and Separation* (mcCS) index for measuring the cohesion and separation of a meta-cluster which is inspired by the popular internal cluster evaluation index - *Silhouette Coefficient* [35], as follows. For a clustering C_i in a meta-cluster $\mathcal{C}_m$, we define

$$mcCS(C_i) = \frac{metaInter(C_i) - metaIntra(C_i)}{\max\{metaInter(C_i), metaIntra(C_i)\}}$$

   where $metaIntra(C_i)$ is the average dissimilarity of clustering C_i with all other clusterings in the same meta-cluster $\mathcal{C}_m$, and $metaInter(C_i)$ is the smallest average dissimilarity of C_i to any other meta-cluster which clustering C_i is not a member. For $metaIntra(C_i)$, the lower the value, the better which indicates the better cohesion. For larger $metaInter(C_i)$, it means the meta-cluster is better separated with others. The dissimilarity between a pair of clusterings, C_i and C_j, is computed by $1 - AMI(C_i, C_j)$. For the meta-cluster $\mathcal{C}_m$, we compute its cohesion and separation by taking the average $mcCS$ of all its member clusterings. For this measure, the larger its value, the more likely the meta-cluster corresponds to a reasonable and distinctive clustering view.

$$mcCS(\mathcal{C}_m) = \frac{1}{|\mathcal{C}_m|} \sum_{C_i \in \mathcal{C}_m} mcCS(C_i) \tag{7}$$

2. Size of the Meta-Cluster
   We also consider the size of each meta-cluster[5]. If the size of the meta-cluster is

[5] Even though, we are not sampling the clustering space uniformly, but the size of the meta-cluster can be considered as one reasonable standard.

large, it indicates that the clustering view is popular according to the available clustering generation methods. Then, we are more confident that this clustering view corresponds to a popular one, since many base clusterings are represented by it. For the meta-clusters with smaller size, it does not necessarily mean that they are not good or not important. It is just that the generation techniques cannot find it easily. The small size can also be taken as one choice but in this paper we choose to prefer large sized meta-clusters.

$$S(\mathcal{C}_m) = \frac{1}{|\mathcal{C}_m|} \tag{8}$$

3. Quality of the Meta-Cluster
We believe that the meta-cluster with better quality of clustering members may result in a clustering view with better quality. We compute the quality of a meta-cluster $\mathcal{C}_m$ as the average quality of its clustering members. For each clustering C_i, we quantify its quality by taking the conditional entropy $H(C_i|X)$ which is same as in the equation 6. The smaller the value is, the more likely that meta-cluster is of high quality.

$$Q(\mathcal{C}_m) = \frac{1}{|\mathcal{C}_m|} \sum_{C_i \in \mathcal{C}_m} H(C_i|X) \tag{9}$$

According to these different criteria, we can get different rankings for the generated meta-clusters. We can choose any of these rankings for returning the top K meta-clusters. In this paper, we rank the meta-clusters in terms of the harmonic mean of the above different rankings, which works reasonable well in our case. But we stress that any combination of the above different measures for ranking can be used [36, 34, 23].

$$averMC(\mathcal{C}_m) = \frac{3}{\frac{1}{RankCS(\mathcal{C}_m)} + \frac{1}{RankS(\mathcal{C}_m)} + \frac{1}{RankQ(\mathcal{C}_m)}} \tag{10}$$

7 Discovering the Clustering Views via Ensemble Clustering

In this final step, we use three ensemble clustering algorithms [37] - CPSA, HGPA and MCLA to find a consensus view for each meta-cluster. Among these three algorithms, MCLA produce the best ensemble clustering in terms of AMI between the generated clustering view and the ground truth clustering view. Thus, we finally choose to present results generated by MCLA ensemble clustering algorithm. However, our framework is not restricted to any specific ensemble clustering method and users can choose the one they prefer according to their requirements. At the end of this step, we have a set of high quality and diverse views of the data.

8 Time Complexity Analysis

In this part, we analyze the time complexity of the proposed rFILTA framework. As rFILTA consists of different steps that involves in different methods, its complexity depends on the time complexities of these different algorithms. Next, we provide detailed analysis of the time complexity for each step in rFILTA.

Time Complexity of Generation Step We employ 7 clustering methods in the generation step (refer to Section 9.1 for details). Time complexities of these methods are: K-means $O(Inkd)$; random feature weighting method $O(Inkd)$; random sampling method $O(Inkd)$; spectral clustering $O(n^3)$; EM clustering $O(Inkd^2)$; information theory based clustering $O(n^2)$; minCEntropy $O(n^2d)$, where I indicates the number of iterations and we set $I = 100$ as default, k indicates the number of clusters in a clustering, d indicates the number of features of a data object and n indicates the number of data objects.

Thus, the overall time complexity of the generation step is $O(Inkd^2+n^2d+n^3)$.

Time Complexity of Filtering Step Before the filtering step, we can pre-compute the meanNN differential entropy $H(X|C_i)$ for all the l base clusterings and also the AMI between each pair of clusterings. In reality, there are only $l(l-1)/2$ pairs of AMI need to be computed due to the symmetry.

In particular, for one clustering C_i, the computation of its meanNN differential entropy $H(X|C_i)$ costs $O(n^2d)$. Thus it costs $O(l \cdot n^2d)$ for all the l clusterings. The entropy of a clustering $H(C_i)$ costs $O(n)$. The mutual information between a pair of clusterings C_i and C_j, $I(C_i, C_j)$, costs $O(n + k^2)$. It costs $O(kn)$ for the computation of the expectation of the mutual information between a pair of clusterings C_i and C_j, $E\{I(C_i, C_j)\}$. Thus, the computation of the AMI of a pairs of clusterings costs about $O(kn)$. Then it costs $O(l^2kn)$ for the computation of AMI for the $l(l-1)/2$ pairs of clusterings. Thus, the pre-computation step costs $O(n^2ld + l^2nk)$. The incremental selection procedure costs $O(L \cdot l)$.

Overall, the time complexity of the filtering step is dominated by the pre-computation step, that is $O(n^2ld + l^2nk)$.

Time Complexity of Grouping Step Please note that after filtering step, we only have L base clusterings left for the following steps. The pairwise dissimilarity matrix has been computed in the filtering step. Using the dissimilarity matrix as input for iVAT algorithm, we can get the reordered dissimilarity matrix as output. This step costs $O(L^2)$. Taking the reordered dissimilarity matrix as input for CLODD algorithm, then we can get the groups of meta-clusters as output and it costs $O(L^3 \cdot N_p \cdot q_{max})$. As CLODD method used particle swarm optimization method [19], N_p is the number of particles for each swarm and q_{max} is the maximum number of swarm iterations.

Time Complexity of Ranking Step In the ranking step, we proposed three criteria for ranking meta-clusters. The time complexities of these three criteria are discussed as follows.

- Cohesion and Separation: As the dissimilarities between each pair of clusterings have been precomputed, for all L clusterings, the complexity of computing mcCS score is $O(LR)$, where R is the number of generated meta-clusters. The worst case is that L meta-clusters generated, i.e., each clustering is a meta-cluster, then it becomes $O(L^2)$. Thus, the overall complexity of this step is $O(L^2)$ in the worst case.
- Quality: As the quality, $H(X|C_i)$, for each clustering C_i has been precomputed, this step costs $O(L)$.
- Size: This step costs $O(L)$.

  Thus, the overall complexity of ranking step is $O(L^2)$.

Complexity of Ensemble Step The time complexity of ensemble step with MCLA is $O(nk^2L^2)$.

In summary, the time complexity of the rFILTA framework might be dominated by the generation step or filtering step or the grouping step depending on the specific datasets. In our case, the grouping step is most expensive in the framework. Please refer to Section 9.6 for more information.

9 Experimental Results

In this section, we evaluate the performance of our rFILTA method against the other existing meta-clustering methods which are all considered as the unfiltered meta-clustering method. We compare their ability to recover known clustering views in 8 datasets (2 synthetic datasets and 6 real world datasets). We also evaluate and compare with alternative clustering method for discovering multiple clustering views. In addition to that, we will show the proposed ranking scheme works well.

Next, we first introduce the clustering methods employed in the generation step. Then we introduce the evaluation scheme for validating the generated multiple clustering views. Then the parameter setting is discussed. Finally we show and analyze the experimental results on different datasets which will demonstrate the performance of our proposed filtering step and ranking step in the rFILTA framework.

9.1 Generation Methods

We employ 7 clustering generation methods in our experiments, many of which have been used previously in other meta-clustering algorithms.

- K-means with random initializations [5].
- Random feature weighting method where feature weights are drawn from the zipf distribution [5].
- Random sampling that selects $\{50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$ of objects and features, and then applying k-means on the sampled objects and features. Then the objects not initially sampled are assigned to the nearest clusters by the k-nearest neighbour method.
- Spectral clustering method [5] using the similarity measure $S = exp(-\|x_i - x_j\|^2/\sigma^2)$ with the shape parameter $\sigma = \frac{\max\{\|x_i - x_j\|\}}{2^{k/8}}$, where k is randomly chosen from $k = 0, \ldots, 64$.
- EM-based mixture model clustering method with different initializations.
- Information theory based clustering algorithm [12].
- An alternative clustering method, minCEntropy [39], with different reference clusterings generated by k-means method.

Generally, we generate 700 base clusterings for each dataset. Each clustering algorithm generates 100 clusterings. The number of clusters in each generated clustering is as same as the ground truth views'.

For the data selection, we would like to keep the number of clusters in both ground truth views consistent. If the number of clusters in ground truth views were

different, then base clusterings with different number of clusters must be generated. This will be challenging for the subsequent grouping and ensemble steps. Research on techniques for grouping and ensemble steps are out of the scope in this work. Our focus is on validating the proposed filtering and ranking methods. Therefore, we choose to keep the number of clusters in the ground truth views consistent and generate base clusterings with same number of clusters, so as to be less distracted by other steps in the framework.

9.2 Evaluation of the Clustering Views

In order to evaluate the goodness of the discovered clustering views, we need to answer the following questions:

- How many ground truth clustering views can be recovered (diversity)?
- How well do the generated clustering views match the multiple sets of ground truth clustering views (quality)?
- How do the returned top K clustering views match the ground truth clustering views (ranking)?

Inspired by *Mean Average Precision* (MAP) [25], a popular measure for evaluating ranked retrieval of documents in information retrieval, we propose a *Mean Best Matching* (MBM) score to evaluate our method. Here, we assess the matching between the returned top K clustering views and the ground truth labels using AMI. In more detail, given multiple ground truth labels $\mathcal{G} = \{G_1, \ldots, G_H\}$ and a set of ranked clustering views $\mathcal{V}_r = \{V_{r_1}, \ldots, V_{r_m}\}$, the MBM for the top K clustering views $\mathcal{V}_{r_k} = \{V_{r_1}, \ldots, V_{r_k}\}$, where $k \leq m$, is defined as:

$$MBM(\mathcal{V}_{r_k}) = \sum_{i=1}^{H} \max_{V_j \in \mathcal{V}_{r_k}} AMI(G_i, V_j)/H \tag{11}$$

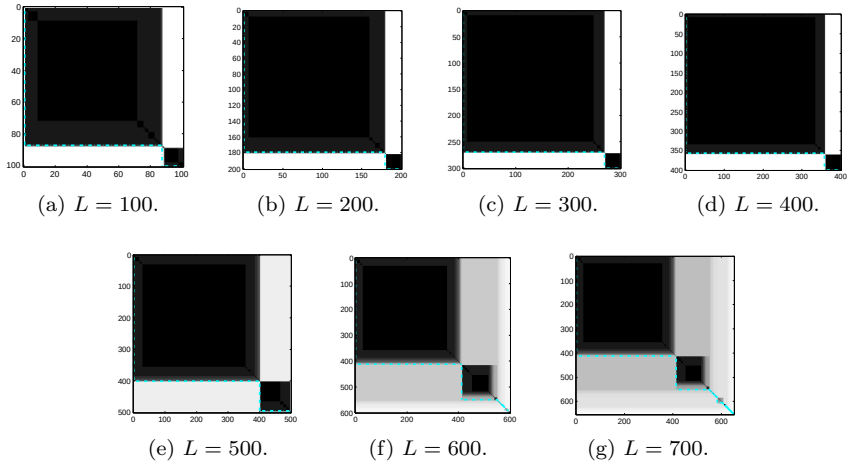
where $MBM(\mathcal{V}_{r_k}) \in [0, 1]$. The $MBM(\mathcal{V}_{r_k})$ takes 0 when there is no ground truth views recovered at all. It takes 1 when all the ground truth views are recovered and matched perfectly by the generated clustering views. The MBM score will increase when the generated clustering views match the ground truth labels better or there is new ground truth views recovered.

9.3 Parameter Setting

In this part, we discuss two important parameters involved in our method, i.e., L and β, the number of selected base clusterings and the regularization parameter used for balancing the quality and diversity during filtering.

9.3.1 The Impact of the Number of Selected Base Clusterings

The number of selected clusterings L does not have high impact on the quality of view generation by our method. We take the flower dataset as the example to show the impact of L. In Figure 5, we show the iVAT diagram constructed for $L = 100$ to 700 (filtered) base clusterings (recall that there are 700 raw base clusterings).



(a) $L = 100$. (b) $L = 200$. (c) $L = 300$. (d) $L = 400$.

(e) $L = 500$. (f) $L = 600$. (g) $L = 700$.

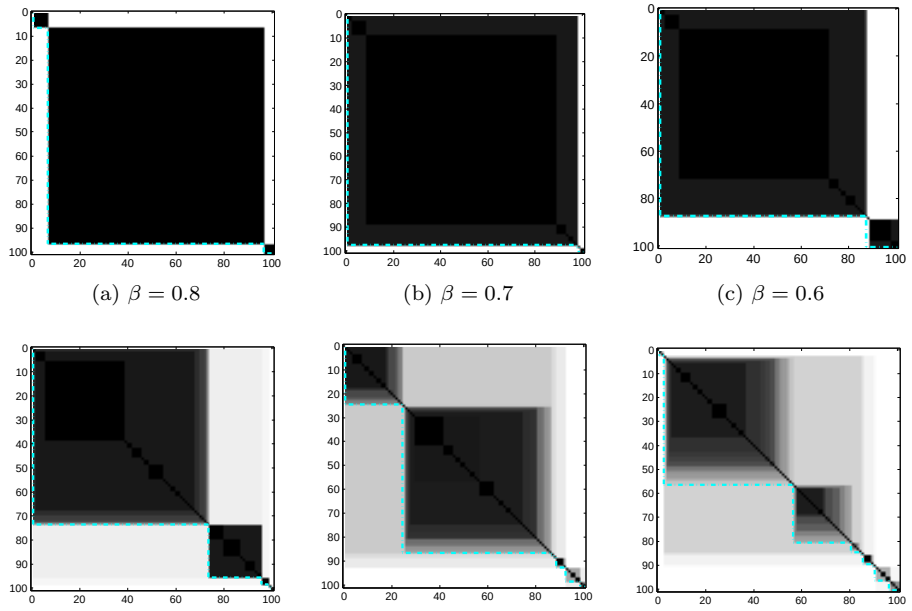
Fig. 5: iVAT diagrams for different number of filtered base clusterings with $\beta = 0.6$ on the flower dataset.

We see that the iVAT diagrams are mostly stable from $L = 100$ to 500, meaning that rFILTA is quite robust to noise and relatively insensitive to the choice of L. We recommend choose $L = 15\% \sim 25\% \times n$. We also obtain similar patterns with the other datasets.

9.3.2 Impact of the Regularization Parameter

The regularization parameter $\beta \in [0, 1]$ balances the quality and diversity during the clustering filtering procedure. For example, when $\beta = 0.5$, it means we treat quality and diversity equally important. When $\beta \to 0$, the filtering process places more emphasis on diversity, which generally increases the number of potential clustering views but at the risk of including more poor quality solutions. In contrast, when $\beta \to 1$, the filtering procedure focuses on the quality, which will result in high quality clustering views but some relevant clustering views may be filtered out. Thus, users can tune this parameter according to their specific needs for view detection. In our experiments, we chose the value for β for each dataset by testing and tuning β within the $[0, 1]$ range according to the intuition about the balance between quality and diversity.

Given that we usually do not have the cluster labels, the iVAT diagrams can be used as one of the ways to help users for investigation. In particular, we propose to 'slide' β within the $[0, 1]$ range and inspect the iVAT reordered matrix and the consensus views that emerge. We take the flower dataset as an example and illustrate a number of iVAT diagrams (Figures 6) constructed from different β values and $L = 100$. We can see that as β decreases, the iVAT diagram becomes more fuzzy, which means that the selected base clusterings are more diverse but their quality is decreasing. Depending on users' different requirements, if they want the dominant and easily identified clustering views, they can put more focus on quality by setting β with large value. If they would like to get more diverse but



(a) $\beta = 0.8$ (b) $\beta = 0.7$ (c) $\beta = 0.6$

(d) $\beta = 0.5$ (e) $\beta = 0.4$ (f) $\beta = 0.3$

Fig. 6: iVAT diagrams generated from 100 filtered base clusterings and different β values, for the flower data.

may be not dominant clustering views, they can put more focus on diversity by tuning β with smaller value. When decreasing β to certain point, the structure and fuzziness of iVAT diagram does not change much. It indicates that diversity might have reached its limit.

Next we will present detailed experimental results and analysis based on 8 datasets, including 2 synthetic datasets and 6 real world datasets.

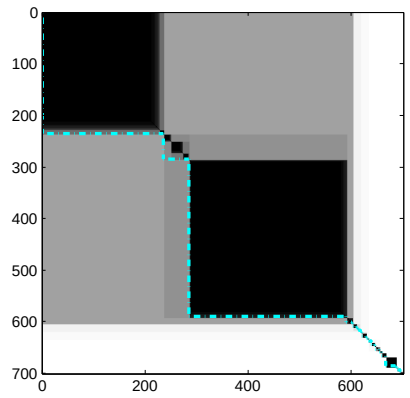
9.4 Synthetic Datasets

In this section, we use 2 synthetic datasets to demonstrate that our rFILTA method is able to discover high quality and diverse clustering views by filtering out poor quality and redundant base clusterings.

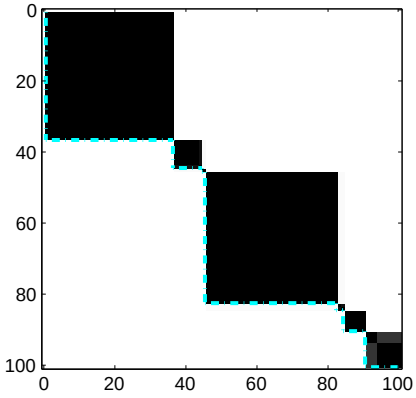
9.4.1 4 Gaussian 2D dataset

The first synthetic dataset is a 2D four Gaussian dataset with 200 data objects (refer to Figure 8a), consisting of 7 ground truth clustering views (refer to Figure 9). We generate 700 base clusterings with 2 clusters.

We first perform meta-clustering on the whole set of generated base clusterings. The iVAT diagram of the unfiltered base clusterings is shown in Figure 7a. We got 30 meta-clusters which are highlighted by the green dashed line surrounding the

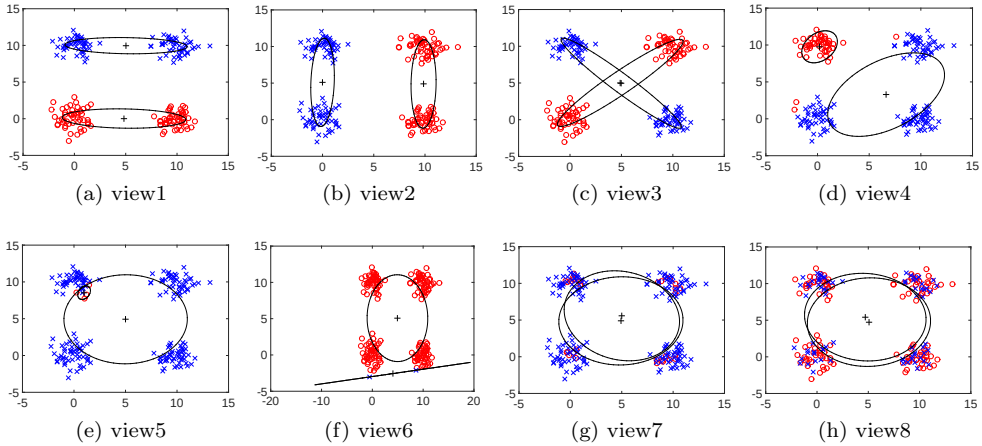


(a) iVAT diagram of the 700 unfiltered base clusterings.



(b) iVAT diagram of the 100 filtered base clusterings.

Fig. 7: iVAT diagrams of the unfiltered and filtered base clusterings on the 2D Gaussian dataset.

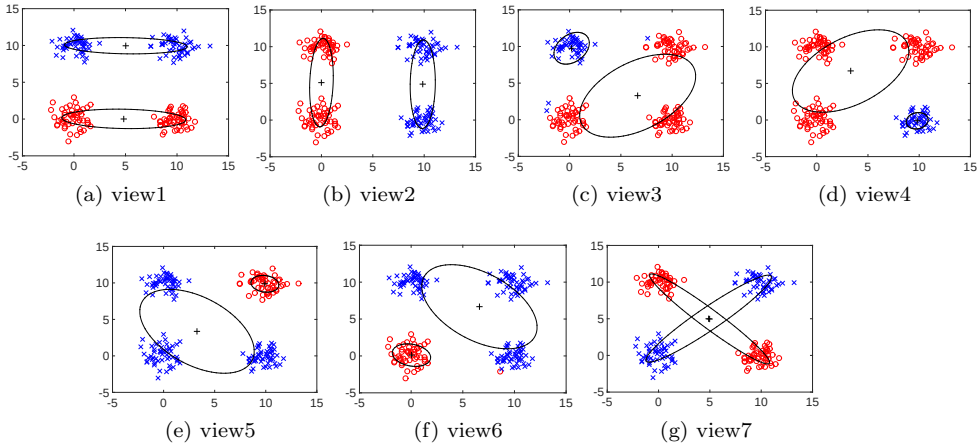


(a) view1 (b) view2 (c) view3 (d) view4

(e) view5 (f) view6 (g) view7 (h) view8

Fig. 8: Top 8 clustering views returned from the unfiltered base clusterings on the 2D Gaussian dataset.

blocks, where each block corresponds to a meta-cluster. Due to the limitation of space, we choose the top 8 meta-clusters, apply ensemble clustering on them and obtain the top 8 clustering views, shown in Figure 8. We can see that the first 4 clustering views are reasonable clustering views, while clustering views 5 to 8 are of poor quality (and also the rest of the clustering views, refer to Figure 10). As we introduced before, this dataset contains 7 ground truth clustering views. The unfiltered meta-clustering method only recovered 4 of them, i.e., the first 4 clustering views shown in Figure 8. Next we apply our filtering approach on the same set of 700 base clusterings. We filter out 600 of the low quality and



(a) view1 (b) view2 (c) view3 (d) view4

(e) view5 (f) view6 (g) view7

Fig. 9: 7 clustering views generated from the filtered base clusterings on the 2D Gaussian dataset.

similar base clusterings setting $L = 100$ and $\beta = 0.6$. The iVAT diagram of the filtered set of base clusterings is shown in Figure 7b consisting of 7 blocks (meta-clusters). The clustering views, generated by applying the ensemble clustering on these meta-clusters, are shown in Figure 9. These 7 clustering views correspond to the 7 ground truth clustering views.

By comparing the results got from the unfiltered and filtered base clusterings, we obtain the following observations:

- *Missing clustering views*: some of the interesting clustering views may be missed while performing meta-clustering on the unfiltered base clusterings. It is because that some meta-clusters are loosely connected by some noisy clusterings and then these meta-clusters are considered as one. Then we will miss some clustering views. Our filtering step can help clean out these noisy clusterings and recover those missed interesting clustering views.
- *Poor quality clustering views*: from the unfiltered base clusterings, we may generate a lot of poor quality clustering views. This is because the generated base clustering may include many poor quality base clusterings which will result in poor quality clustering views (refer to Figure 7a and Figure 8). Our filtering step can help clean out these poor quality base clusterings, and present those good quality ones (refer to Figure 7b and Figure 9).

To obtain further insights, let us examine the MBM scores for these two sets of clustering views shown in Figure 10. The x axis indicates the value of K for the returned top K clustering views in terms of our average ranking scheme, e.g., $K = 4$ indicates the returned top 4 clustering views. The y axis shows the corresponding MBM scores. The blue crosses indicate the results from the unfiltered base clusterings, and the red circles indicate the results from the filtered base clusterings.

1. There are 30 clustering views generated from the unfiltered base clusterings. There are 7 clustering views generated from the filtered base clusterings with setting $L = 100, \beta = 0.8$.
2. The MBM scores for the unfiltered base clusterings are increasing up to $K = 4$ clustering views. The increasing of MBM scores may be due to a new ground truth clustering view being recovered or the newly returned kth clustering view matching one of the recovered ground truth views better. Here, it is the first case. Then, the MBM scores plateau for $K \geq 5$ clustering views. When MBM scores do not change with an increasing K, if it has not got the best matching of all the ground truth clustering views, i.e., $MBM(\mathcal{C}_m) = 1$, then it means that either no new ground truth clustering view is discovered, or the newly returned kth clustering views do not match better with the discovered ground truth clustering views for returned first $k - 1$ clustering views. Here, it is both. Thus, with the unfiltered base clusterings, we discovered 4 ground truth clustering views.
3. The MBM scores for returned top $K = 7$, clustering views from the filtered set of base clusterings are increasing. It is because it recovers a new ground truth clustering view each time. Finally, it recovers all the 7 ground truth clustering views and reach an almost perfect matching with MBM score close to 1.

9.4.2 8 *Gaussian 3D dataset*

Next, we show the experimental results on a 3D synthetic dataset with 800 data objects which contains 8 Gaussian clusters. We generate 700 base clusterings with 2 clusters on this dataset. There are 3 ground truth clustering views for this dataset (refer to Figure 13).

The iVAT diagram of the unfiltered base clusterings is shown in Figure 11a. We discovered 52 clustering views, corresponding to the 52 meta-clusters presented as diagonal blocks in the iVAT diagram, from this set of unfiltered base clusterings. We show the top 4 clustering views in Figure 12. The first 3 clustering views are corresponding to the three ground truth clustering views. However, the fourth clustering view is of poor quality.

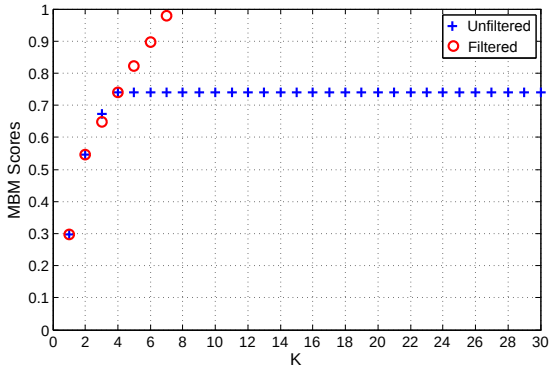
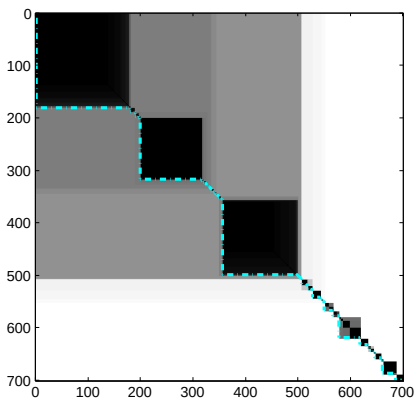
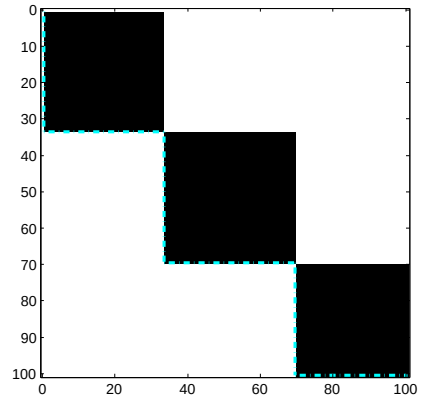


Fig. 10: MBM scores of the clustering views generated from the unfiltered and filtered base clusterings on the 2D Gaussian dataset.



(a) iVAT diagram of the 700 base clusterings.



(b) iVAT diagram of the 100 filtered base clusterings.

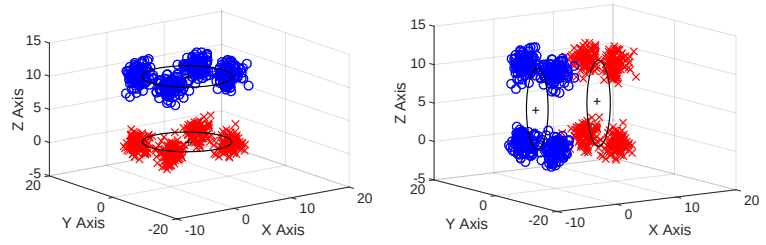
Fig. 11: iVAT diagrams of the unfiltered and filtered base clusterings on the 3D Gaussian dataset.

From this set of 700 base clusterings, we keep 100 base clusterings after filtering with $\beta = 0.7$. The iVAT diagram of the filtered base clusterings is shown in Figure 11b. We can see that there are three clearly separated blocks which correspond to the three ground truth clustering views (Figure 13).

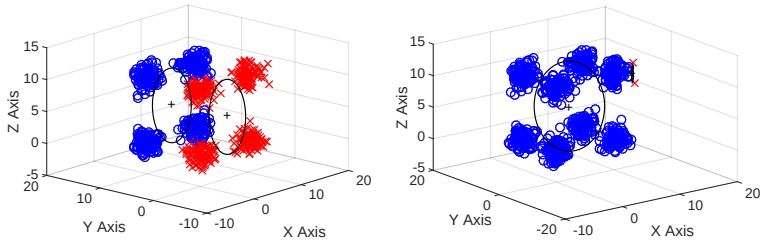
The MBM scores for these two sets of clustering views are shown in Figure 14. We can observe the following:

1. There are 52 clustering views generated from the unfiltered base clusterings. After filtering with $L = 100, \beta = 0.7$, we discovered 3 clustering views.
2. The top 3 clustering views returned from the unfiltered base clusterings recover the three ground truth clustering views and match the ground truth clustering views perfectly with $MBM(\mathcal{C}_3) = 1$. The MBM scores are invariant after the 3rd clustering view. It is because the returned first 3 clustering views have recovered all the 3 ground truth views and matched them perfectly. For the other clustering views, we inspect that they contain a lot of poor quality and redundant ones which is due to the redundant and poor quality base clusterings.
3. The 3 clustering views got from the filtered set of base clusterings recover and match the three ground truth clustering views perfectly.

In this set of experiments, we find that we can recover the 3 ground truth clustering views perfectly from the unfiltered base clusterings. Do we still need filtering? We can observe that after filtering, we can get rid of the irrelevant clustering views by filtering out the redundant and poor quality clusterings, and just obtain the 3 ground truth clustering views in this case. In addition to that, for the case of 'missing clustering views' found out in the experiments on the 4 Gaussian dataset, only ranking does not help solve this problem. Thus, the filtering step is necessary. We will further discuss the necessity of filtering in the following experiments.

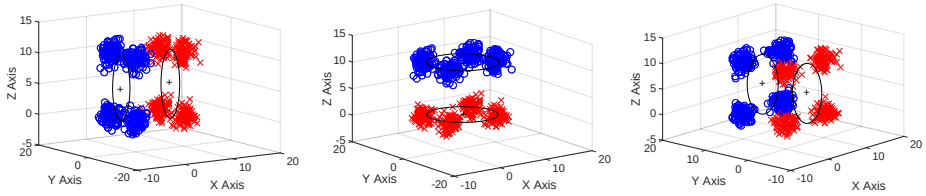


(a) View1 (b) View2



(c) View3 (d) View4

Fig. 12: The top 4 clustering views discovered from the 700 unfiltered base clusterings on the 3D Gaussian dataset.



(a) View1 (b) View2 (c) View3

Fig. 13: The 3 clustering views discovered from the 100 filtered base clusterings on the 3D Gaussian dataset.

9.5 Real Datasets

In the following section, we will evaluate rFILTA on 6 real datasets. These datasets cover a variety of possible datasets.

9.5.1 CMUFace Dataset

The CMUFace dataset from the UCI Machine Learning Repository [2] is a commonly used dataset for the discovery of alternative clusterings [6]. It contains 624, 32×30 pixels images of 20 persons, along with different features of these persons, e.g., pose (straight, left, right, up). Two dominant clustering views exist in this dataset - person (identity) and pose. In our experiment, we randomly select the

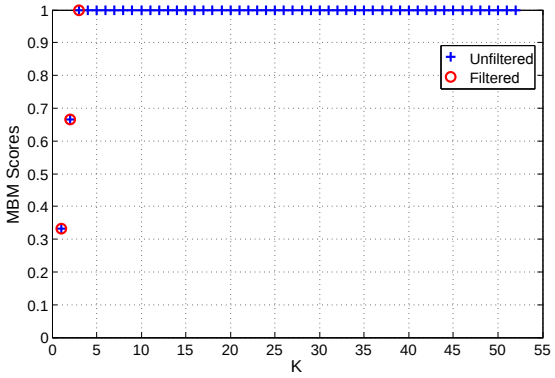


Fig. 14: The MBM scores of the clustering views generated from the unfiltered and filtered base clusterings on the 3D Gaussian dataset.

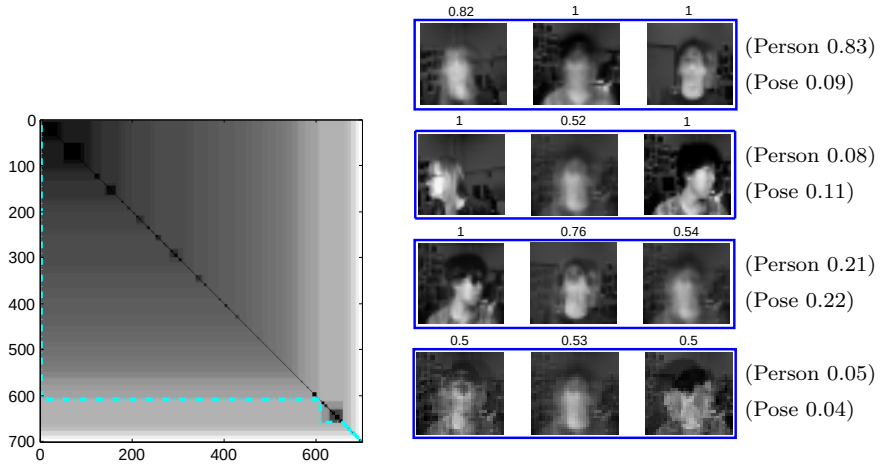
Fig. 15: Two ground truth clustering views on CMUFace dataset. The first row is person view and the second row is pose view.

images of three people and have 93 images in total. Again we generated 700 base clusterings and rFILTA selected $L = 100$ base clusterings. The two ground truth clustering views are shown in Figure 15. Each image is represented as the mean of images within the cluster.

Next, we will show three sets of experimental results on this dataset to show the benefits of filtering and ranking.

Benefits of filtering

In this experiment, we demonstrate the necessity of filtering. For the 700 unfiltered base clusterings, their iVAT diagram is shown in Figure 16a. We found 23 clustering views from this set of unfiltered base clusterings. Due to the limitation of space, we show the top 4 clustering views in Figure 16b. Each row represents a clustering view and each clustering view consists of three clusters, which are shown as the mean of all the images in each cluster. The number above each image is the percentage



(a) The iVAT diagram of the unfiltered base clusterings.

(b) The top 4 clustering views from the unfiltered base clusterings. The score above each image is the percentage of the dominant class in that cluster. The AMI similarity scores between each clustering view with the two ground truth views respectively are presented at the right side of each clustering view.

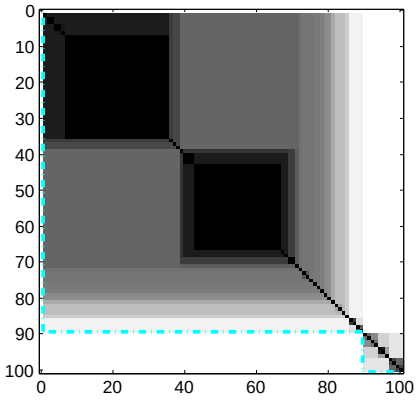
Fig. 16: Results for the unfiltered base clusterings on CMUFace dataset.

of the dominant class in this cluster and indicates the purity of this cluster (it is the same in the later experiments. For simplicity, we will not explain it again). As we can see from these four clustering views, the first row is the person view. However, the pose view is not presented in the other three, nor in the other 19 clustering views. It is because that the pose clustering view is hidden among the other meta-clusters due to the noisiness of the base clusterings. From this set of experiments, we find out that we may miss out some interesting clustering views from the unfiltered base clusterings due to the noisiness of the generated base clusterings.

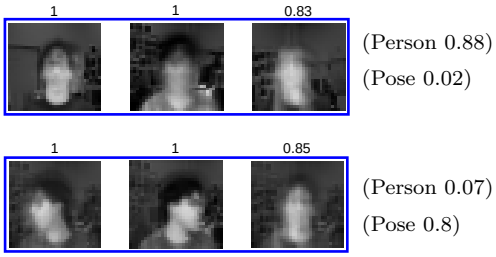
Next, we show the results on the 100 filtered base clusterings with $\beta = 0.6$ in Figure 17. As we can see the iVAT diagram of the filtered base clusterings in Figure 17a, it contains two clearly separated blocks. Examining the clustering views obtained from these two blocks shown in Figure 17b, they are exactly the person and pose views that we are looking for. Compared with the results shown in Figure 16, we can see that after filtering, the resulting iVAT diagram is less noisy, and the blocks are more clear and well separated. After filtering out the noisy base clusterings, we have recovered the hidden pose views.

Benefits of ranking

When we got multiple clustering views, particularly the number of clustering views is large, it is time consuming to examine them all. Our filtering step can help reduce



(a) iVAT diagram of the filtered base clusterings with $\beta = 0.6$.

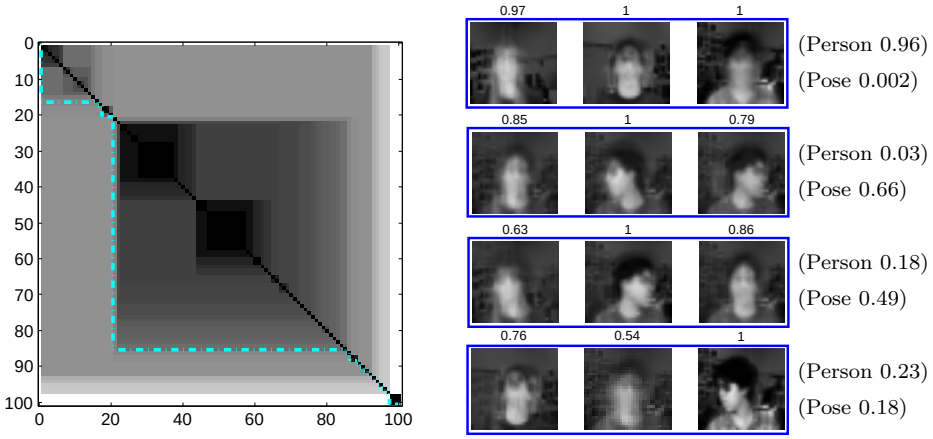


(b) The 2 clustering views generated from the filtered base clusterings. The score above each image is the percentage of the dominant class in that cluster. The AMI similarity scores between each clustering view with the two ground truth views respectively are presented at the right side of each clustering view.

Fig. 17: Results for the filtered base clusterings on CMUFace dataset.

the number of clustering views by filtering out the poor quality and similar ones, and discover the good clustering views (Figure 17). However, sometimes there are still a lot of potentially interesting clustering views after filtering, especially when we do not know how many interesting clustering views exist. It depends on different factors, e.g., the complexity of datasets, the distribution of the generated base clusterings, and different requirements of users. Hence, it will be helpful to rank these clustering views and make it easier for users to analyze them.

For example, users may want to explore more potentially interesting clustering views exist in the generated base clusterings by adjusting the tradeoff parameter β. When the value of β is large, we may get a few good and diverse clustering views (e.g., Figure 17). When they decrease the value of β, the filtered set of base clusterings will be more diverse which may result in more clustering views. For the same set of 700 base clusterings, when we decrease $\beta = 0.4$, we discovered 16 clustering views. Refer to the Figure 18a, the iVAT diagram contains more blocks and is more fuzzy compared with the iVAT diagram with $\beta = 0.6$ shown in Figure 17a. It is not easy to examine all of them. Thus, it will be helpful if we can rank these clustering views and recommend the top K to users for facilitating their job. As shown in Figure 18b, it is the returned top 4 clustering views by our ensemble ranking scheme. The first row is the person view, and the second row is pose view. In this way, users can check less but more possible interesting clustering views from the top K ones instead of all.



(a) The iVAT diagram of 100 filtered base clusterings with $\beta = 0.4$.

(b) The top 4 clustering views generated from the filtered base clusterings with $\beta = 0.4$.

Fig. 18: Results for the filtered set of base clusterings on CMUFace dataset.

Benefits of filtering + ranking

Can we directly rank the clustering views generated from the unfiltered base clusterings without filtering? There is one problem with this. When the generated base clusterings are noisy, some of the potential interesting clustering views could not be discovered from them. In this way, ranking does not help to get those missed clustering views. Thus, we need filtering which can help discover potential interesting clustering views hidden in the generated base clusterings. Based on that, the ranking step could facilitate the analysis of these multiple clustering views.

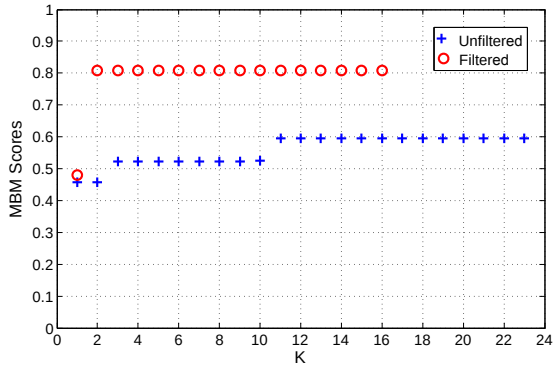


Fig. 19: The MBM scores for the two sets of clusterings views on CMUFace dataset.

The MBM scores for the two sets of discovered clustering views, i.e., filtering with $\beta = 0.4$ and unfiltered base clusterings are shown in Figure 19. In summary:

1. We discovered 23 clustering views from the unfiltered base clusterings and found out 16 clustering views from the filtered base clusterings with $L = 100, \beta = 0.4$.
2. For the filtered base clusterings, the discovered clustering views achieve better MBM scores than the unfiltered base clusterings. It is because after filtering, the two ground truth clustering views are discovered. And they match well with the ground truth clustering views. From the unfiltered base clusterings, we only discover one of the ground truth views.
3. The returned top 2 clustering views from the filtered base clusterings are corresponding to the two ground truth clustering views.
4. Without filtering, only one clustering view can be discovered, even when ranked. Without ranking, the two good and diverse clustering views which are corresponding to the two ground truth views, are hidden among the 16 generated clustering views.

9.5.2 Card Dataset

The card dataset[6] consists of 52 images of cards. It can be explained from different perspectives. A deck of cards can be clustered in terms of different suits (heart, cube, diamond and spade), different colors (red, black, and mixed color), and different rankings (1∼13). In our experiments, we randomly choose three different suits along with all the cards belong to these suits. Finally, we got 39 cards in total containing two clustering views, i.e., suits (spades, diamond and heart) and color (red, black and mixture). We scaled these images to 100×140 pixels. The features of images are described using the HOG descriptors[7] with 2×2 cells [7]. We further applied the *Principle Component Analysis* (PCA) to reduce the number of features to 18, which retains more than 90% variance of the original data. The two ground truth clustering views are shown in Figure 20. Each image is represented as the mean of images within the cluster.

In this set of experiments, we would like to discuss and demonstrate two problems. Firstly, we investigate and discuss the influence of alternative clustering methods in our framework. In addition to that, we show the performance of our filtering and ranking functions.

Alternative Clustering on Card Dataset

In our rFILTA framework, we take alternative clustering method as one of the generation methods for generating diverse base clusterings. As we discussed in the introduction, the alternative clustering algorithms restrict the definition of alternative clustering to certain type of objective functions. When the definition of the alternative clustering is not suited for the clustering structure underlying the data, this approach may not find the alternative clusterings. In this set of experiments,

[6] The images of card are downloaded from https://code.google.com/p/vectorized-playing-cards/.

[7] The code of feature extraction is available at https://github.com/adikhosla/feature-extraction.

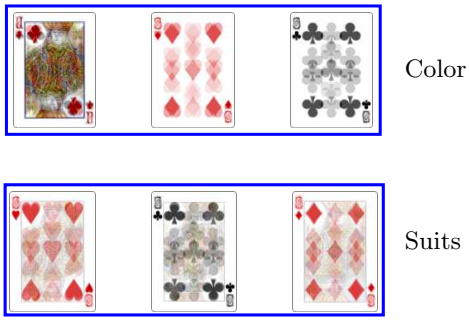


Fig. 20: Two ground truth clustering views on Card dataset. The first row is the color view and the second row is the suits view.

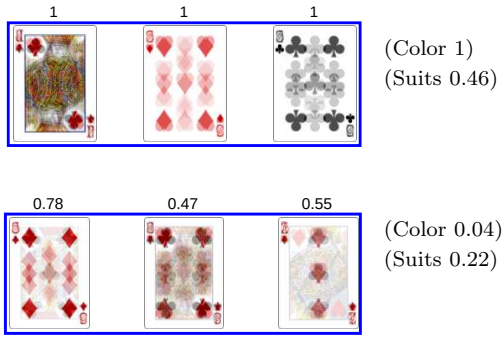
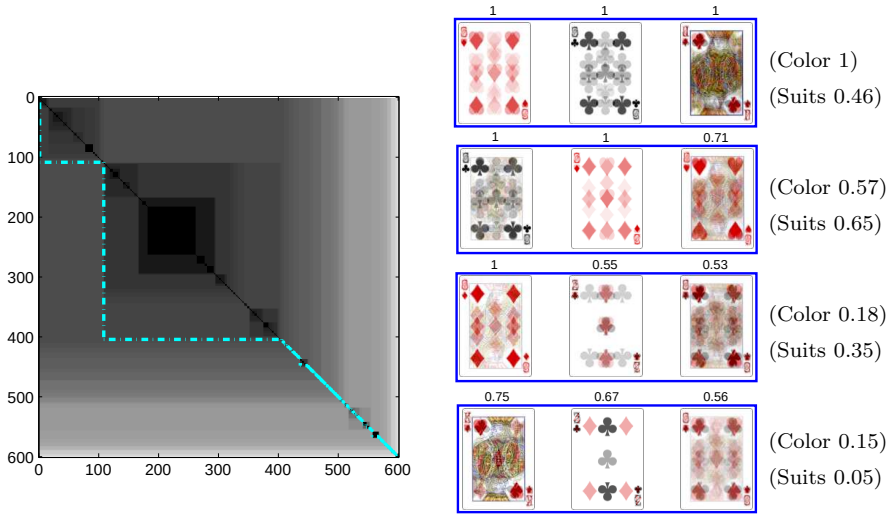


Fig. 21: The alternative clusterings generated by minCEntropy. The first row is the color view given suits view as reference clustering. The second view is generated given color view as reference clustering. The score above each image is the percentage of the dominant class in that cluster. The AMI similarity scores between each clustering view with the two ground truth views respectively are presented at the right side of each clustering view.

we would like to compare the performance of the alternative clustering method (minCEntropy) and rFILTA (without alternative clustering as generation method) on the card dataset.

We apply alternative clustering algorithm, minCEntropy, on the card dataset to generate alternative clusterings. We take one of the two ground truth clustering views (i.e., color and suits) as reference clustering to find the other one. We use the default parameter setting of minCEntropy. The results are shown in Figure 21. The first row is the color view taking the suits view as the reference clustering. This alternative clustering is discovered successfully. The second row is an alternative clustering view generated by taking the color view as given clustering. However, it is not suits view or other explainable view. The possible reason may be that



(a) The iVAT diagram of the 600 unfiltered base clusterings on the card dataset.

(b) The top 4 views generated from the unfiltered base clusterings on the card dataset.

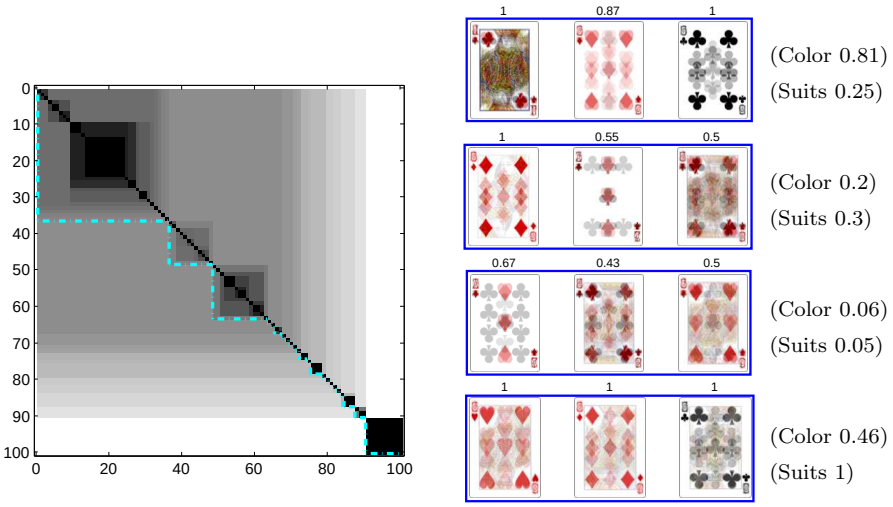
Fig. 22: Results on the 600 unfiltered base clusterings on the card dataset.

the definition of the alternative clustering in this algorithm does not capture the structure of the suits view. Hence, it can be seen alternative clustering can fail to find interesting clustering views if the definition of the alternative clustering does not characterize the alternative clustering properly.

Unfiltered Meta-Clustering without Alternative Clustering Generation on Card Dataset

In this set of experiments, we generate 600 base clusterings with the 6 of the available clustering generation methods without using the alternative clustering method, i.e., minCEntropy. We first do meta-clustering on the whole set of 600 base clusterings. The results are shown in Figure 22. As we can see from the iVAT diagram in Figure 22a, the diagonal blocks are not clearly and well separated. We discovered 135 clustering views from this unfiltered base clusterings. Note that without rankings, this is a very large number to evaluate over. The top 4 clustering views are shown in Figure 22b. The first row is the color view, and the second row is the suits view. In the color view, the three clusters indicate red, black and mixed color respectively from left to right. In the suits view, the three clusters are corresponding to spades, diamond and heart these three suits respectively from left to right.

Comparing the results of the above two sets of experiments, i.e., alternative clustering and meta-clustering, we can observe that the suits view is discovered by the meta-clustering while is not by alternative clustering (Figure 21). This means that the suits view could be captured by some of the clustering methods among the 6 generation methods while it could not be captured by the definition of the alternative clustering in minCEntropy algorithm.



(a) The iVAT diagram of the 100 filtered base clusterings on the card dataset with $\beta = 0.3$.

(b) The top 4 clustering views generated from the 100 filtered base clusterings on the card dataset.

Fig. 23: Results for the filtered base clusterings on the card dataset.

rFILTA without Alternative Clustering Generation on Card Dataset

In this set of experiments, we try to demonstrate the performance of our filtering and ranking function in the rFILTA framework. We apply rFILTA method on the same 600 base clusterings as used in the previous experiments. After filtering, we got 100 base clusterings with $\beta = 0.3$. The results are shown in Figure 23. The iVAT diagram of the 100 filtered base clusterings is shown in Figure 23a. Compared with the unfiltered one (Figure 22a), we can tell that after filtering, some of the meta-clusters (presented as dark blocks along the diagonal) are more clearly presented than before filtering. The top 4 clustering views are shown in Figure 23b. The first row is the color view and the fourth row is the suits view.

The MBM scores for the two sets of clustering views generated from the unfiltered and filtered sets of base clusterings are shown in Figure 24. In summary:

1. We discovered 135 clustering views from the unfiltered base clusterings and 23 clustering views from the filtered base clusterings with $L = 100, \beta = 0.3$.
2. The top 2 clustering views from the unfiltered base clusterings recover the two ground truth views. The top 4 clustering views from the filtered base clustering recover the two ground truth views.
3. The best MBM scores of the clustering views from the filtered set of base clusterings are better than the ones from the unfiltered set of base clusterings. It means that the quality of the recovered clustering views from the filtered set of base clusterings are better than the ones got from the unfiltered set of base clusterings.

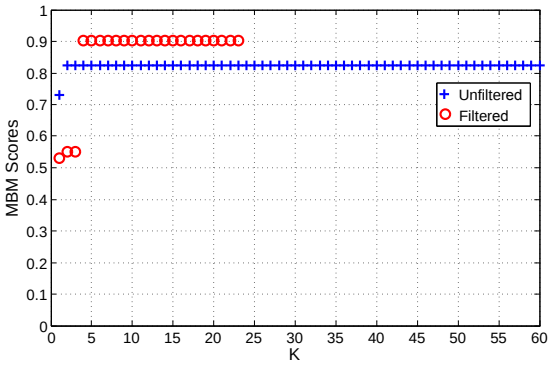


Fig. 24: The MBM scores for the two sets of clustering views from the unfiltered and filtered sets of base clusterings on the card dataset.

9.5.3 Isolet Dataset

The isolet dataset from UCI machine learning repository [2] contains 7797 records with 617 features, which come from 150 subjects speaking the name of each letter of the alphabet twice. There are two clustering views (speaker and letters) in this dataset. In our experiment, we randomly selected 10 persons along with 10 letters, resulting in a 200 records dataset. We generate 700 base clusterings that contains the speaker and letter views, and select 100 base clusterings using rFILTA ($\beta = 0.6$).

The results are shown in Figure 25. Compared with the iVAT diagram of the unfiltered base clusterings in Figure 25a, the iVAT diagram of filtered base clusterings in Figure 25b contains more clear blocks. It may because the clustering views are more easily identified after filtering out the irrelevant base clusterings. The MBM scores of these two sets of clustering views are shown in Figure 25c. In summary:

1. We discovered 92 clustering views from the unfiltered base clusterings and discovered 28 clustering views from the filtered set of base clusterings with $L = 100, \beta = 0.6$.
2. The best MBM scores of filtered clustering views are a little higher than the ones of unfiltered clusterings views. It is because the quality of the clustering views generated from the filtered base clusterings are better than the ones from the unfiltered base clusterings.

9.5.4 WebKB Dataset

The WebKB dataset [8] contains webpages collected mainly from four universities: Cornell, Texas, Washington and Wisconsin. We selected all documents from those four universities that fall under one of four page types namely, course, faculty, project and student. We preprocessed the data by removing common words and rare words (appeared less than twice), stemming. Finally, we choose 350 words

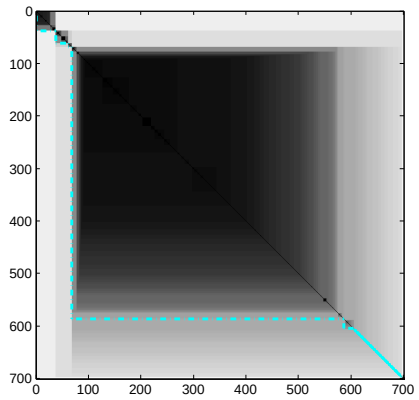
[8] www.cs.cmu.edu/ webkb

with the highest variance. We use the TF-IDF weighting to construct the feature vectors. The final data matrix contains 1041 documents and 350 words. This dataset can be either clustered by the four universities or by the four page types.

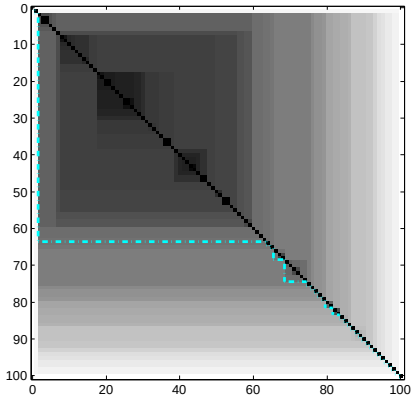
We generated 700 base clusterings and selected 100 base clusterings with $\beta = 0.8$. The results are shown in Figure 26. Comparing the iVAT diagrams from the unfiltered and filtered base clusterings, the iVAT diagram of the filtered base clusterings in Figure 26b reveal more clear blocks while the iVAT diagram from the unfiltered base clusterings in Figure 26a are fuzzy and not clearly separated.

The MBM scores of these two sets of clustering views are shown in Figure 26c.

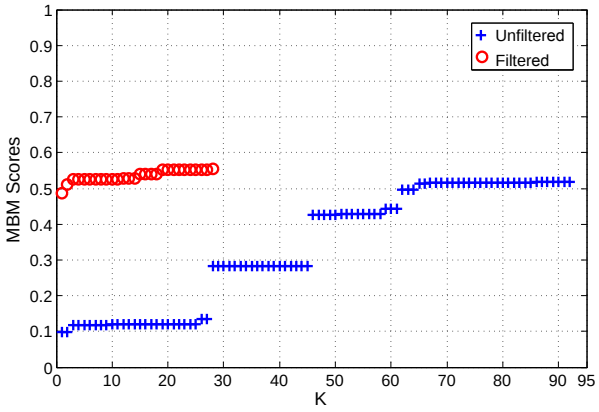
1. We generated 112 clustering views from the unfiltered base clusterings and generated 22 clustering views from the filtered set of base clusterings with $L = 100, \beta = 0.8$.



(a) iVAT diagram of the 700 unfiltered base clusterings.

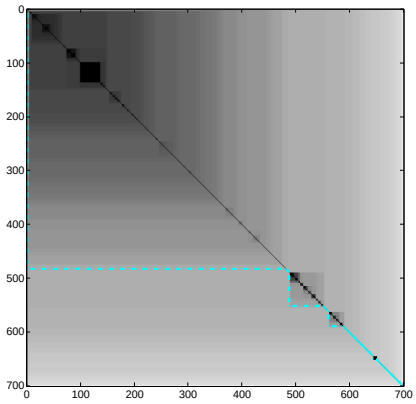


(b) iVAT diagram of the 100 filtered base clusterings with $\beta = 0.6$.

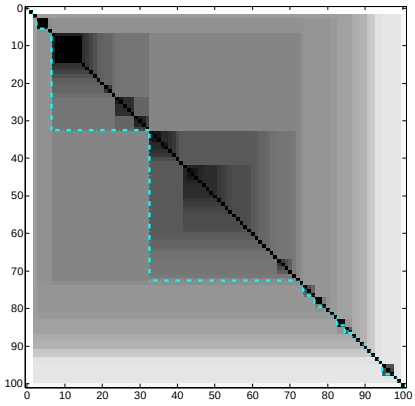


(c) The MBM scores for the clustering views from the unfiltered and filtered set of base clusterings.

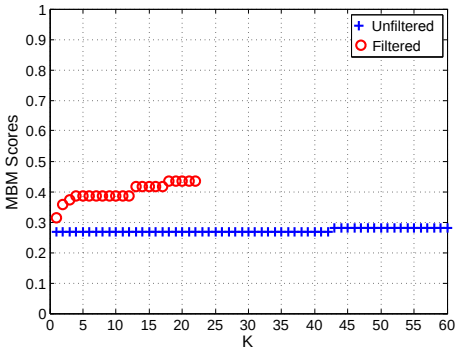
Fig. 25: Results on the isolet dataset.



(a) iVAT diagram of the 700 raw base clusterings.



(b) iVAT diagram of the 100 filtered base clusterings with $\beta = 0.8$.



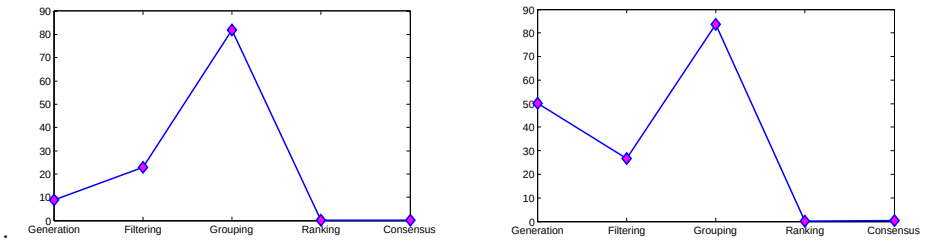
(c) The MBM scores for the clustering views from the unfiltered and filtered base clusterings.

Fig. 26: Results on the webkb dataset.

2. The clustering views generated from the filtered base clusterings are matching the ground truth views better than the ones generated from unfiltered base clusterings.

9.6 Evaluation of Running Time for Each step in rFILTA

In this set of experiments, we show the running time of different steps in rFILTA framework on CMUFace dataset and isolet dataset in Figure 27. As we see from the Figure 27a and Figure 27b, the grouping step takes much more time compared with other steps. It is because the CLODD method used in the grouping step is a genetic algorithm which is slow. Then, the generation step in isolet dataset takes more time than that in the CMUFace dataset. It is because the size of the isolet dataset and the features of the isolet dataset is larger than the CMUFace



(a) The running time of different steps in rFILTA on CMUFace dataset

(b) The running time of different steps in rFILTA on isolet dataset.

Fig. 27: Running time of different steps in rFILTA on CMUFace dataset and isolet dataset in seconds.

dataset. As grouping is not a contribution of this paper, we leave it to future work to explore faster alternative for grouping.

More experimental results and analysis on flower dataset and object dataset are presented in Appendixes Section A and Section B.

10 Conclusions

Meta-clustering is an important tool for discovering multiple views from data by analyzing a large set of raw base clusterings. It does not require any prior knowledge nor pose any assumption on the data, which especially suits exploratory data analysis. However, the generation of a large set of high-quality base clusterings is a challenging problem. There may exist poor quality and similar solutions which will affect the generation of high quality and diverse views.

In this paper we have introduced a clustering selection method for filtering out the poor quality and redundant clusterings from a set of raw base clusterings. This has the effect of lifting the quality of clustering views generated by the meta-clustering methods applied to this set of filtered clusterings. In particular, we proposed a mutual information based filtering criterion which considers the quality and diversity of clusterings simultaneously. By optimizing this objective function via a simple incremental procedure, we can select a subset of good and diverse base clusterings. Meta-clustering on this filtered set of base clusterings can then yield multiple good and diverse views. In addition, we proposed scheme to rank multiple clusterings. We demonstrated that ranking is important when the number of potentially interesting clustering views is large. We believe rFILTA is a simple and useful tool in the area of multiple clustering exploration and analysis.

A Flower Dataset

The flower image dataset [30] consists of 17 species of flowers with 80 images of each. We choose images from 4 species which are Buttercup, Daisy, Windflower and Sunflower (Figure 28). For each specie, we randomly choose 16 images, that is 64 images in total. There are two natural clustering views in this dataset: color (white and yellow) and shape (sharp and round). For reducing the disturbance and focusing on the flowers, we processed the images by blacking the

Fig. 28: Example images of buttercup, sunflower, windflower and daisy flowers from left to right, from top to bottom.

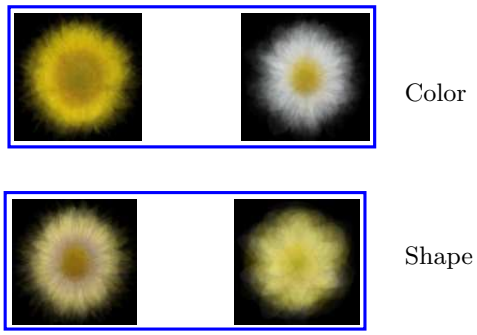
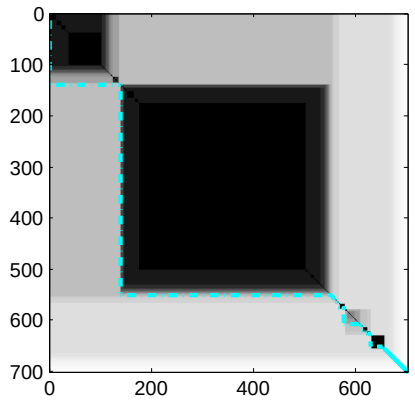


Fig. 29: Two ground truth clustering views on flower dataset. The first row is the color view and the second row is the shape view.

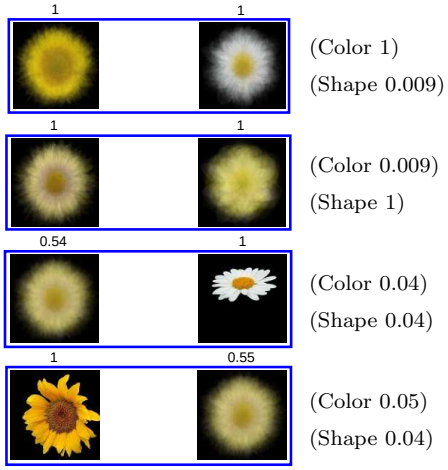
background. We scaled these images to 120×120 pixels and extracted their features in the same way as we did for card dataset. Finally each image is represented by 22 features. We generate 700 base clusterings on this dataset with 2 clusters. The two ground truth clustering views are shown in Figure 29.

We firstly show the results on the unfiltered base clusterings in Figure 30. We got 62 clustering views from this set of unfiltered base clusterings. The top 4 clustering views are shown in Figure 30b. The first row is the color view, containing two clusters, representing two colors, yellow and white. The second row is the shape view, including two shapes, sharp and round. The results after filtering with $L = 100, \beta = 0.6$ are shown in Figure 31. As we can see from the Figure 31a, the iVAT diagram contains two clearly separated blocks (meta-clusters) after filtering out the irrelevant clusterings (compared with unfiltered iVAT diagram in Figure 30a). The generated clustering views from these two meta-clusters are shown in Figure 31b which are just the color and shape view.

To further demonstrate the utility of ranking, we show another set of results in Figure 32 with $L = 100, \beta = 0.3$. When we decrease the tradeoff parameter to $\beta = 0.3$, more diverse clusterings will be included. Thus, the iVAT diagram in Figure 32a is more fuzzy and untidy compared with the higher $\beta = 0.6$ in Figure 31a. We generated 9 clustering views from this set

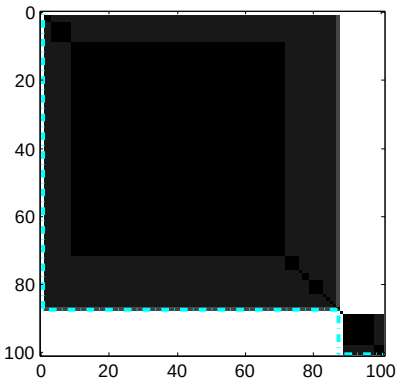


(a) The iVAT diagram of the 700 unfiltered base clusterings on the flower dataset.

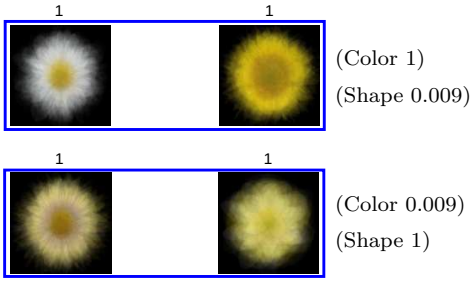


(b) The top 4 views got from the 700 unfiltered base clusterings. The first row is the color view and the second row is the shape view.

Fig. 30: Results for the unfiltered base clusterings on the flower dataset.



(a) The iVAT diagram of 100 filtered base clusterings with $beta = 0.6$ on the flower dataset.

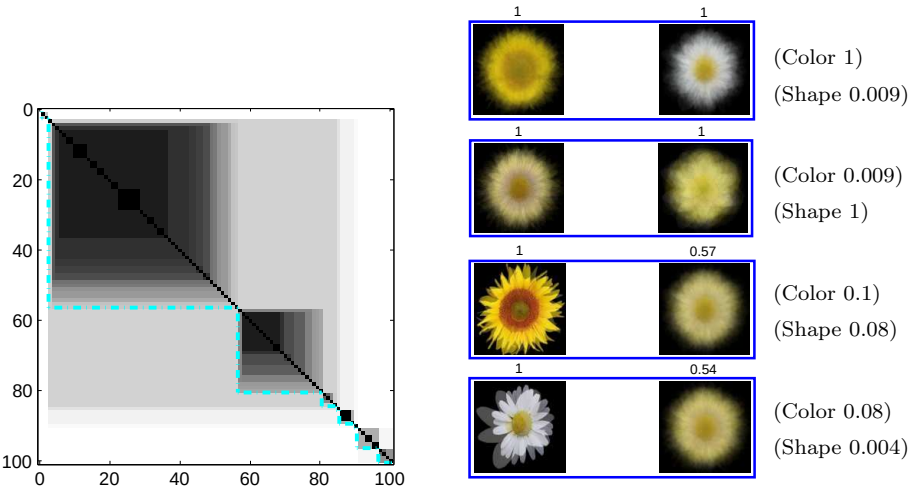


(b) The 2 clustering views got from the filtered base clusterings on the flower dataset. The first row is the color view and the second row is the shape view.

Fig. 31: Results for the 100 filtered base clusterings on the flower dataset.

of filtered set of base clusterings. The top 4 clustering views are shown in Figure 32b. As we can see, the first row is the color view and the second row is the shape view. As we decreased

the tradeoff parameter β, we select more diverse clusterings which result in more clustering views.



(a) The iVAT diagram of the 100 filtered base clusterings with $beta = 0.3$ on the flower dataset.

(b) The top 4 views generated from the filtered base clusterings. The first row is the color view and the second row is the shape view.

Fig. 32: Results of the filtered base clusterings on flower dataset.

The MBM scores for clustering views generated from the unfiltered base clusterings and the filtered base clusterings with $\beta = 0.3$ are shown in Figure 33. In summary:

1. We generate 62 clustering views from the unfiltered base clusterings and generated 9 clustering views from the filtered base clusterings with $\beta = 0.3$.
2. The top 2 clustering views from both sets of clusterings recover and match well with the ground truth clustering views.
3. The rank function works well by ranking the color and shape views as the top 2.

B Object Dataset

The Amsterdam Library of Object Images (ALOI) consists of 110250 images of 1000 common objects. For each object, a number of photos are taken from different angles and under various lighting conditions. We choose 9 objects with different colors and shapes, for a total of 108 images (Figure 34). We processed them in the same way as the card dataset and extracted 15 features for each image finally.

In this set of experiments, we generate 700 base clusterings with 3 clusters. We would like to demonstrate the performance of the filtering and ranking functions in our rFILTA framework. The experimental results on the unfiltered set of base clusterings are shown in Figure 36. As we can see from the iVAT diagram of the 700 base clusterings in Figure 36a, there are a big block and two small blocks along the diagonal. We finally generate 3 clustering views shown in Figure 36b. The first row is the color view, containing three clusters, red, green and yellow. We do not find the shape view from the unfiltered base clusterings. Next, we show the results on the filtered set of base clusterings with $L = 100, \beta = 0.95$ in Figure 37. The iVAT diagram

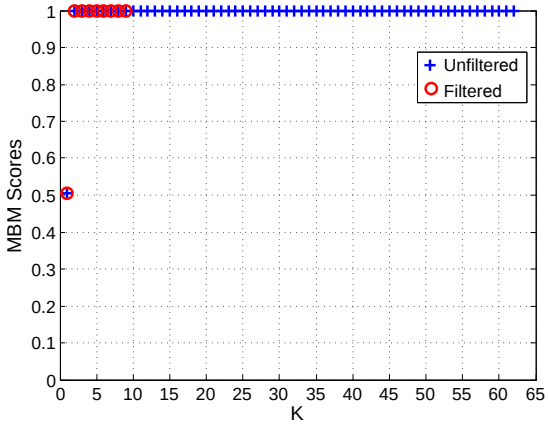


Fig. 33: The MBM scores for two sets of clustering views generated from the unfiltered and filtered base clusterings on the flower dataset.

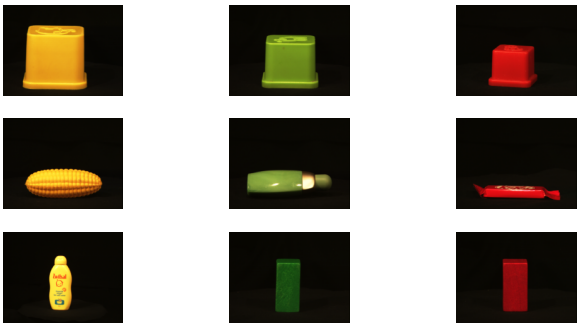


Fig. 34: Example images of the nine selected objects.

contains multiple clear blocks. The top 4 clustering views are shown in Figure 37b. The first row is the color view and the fourth row is the shape view.

Comparing the two sets of results from the unfiltered base clusterings and the filtered base clusterings, we have some observations. There are less clustering views generated from the unfiltered base clusterings than ones from the filtered base clusterings. It may be because that there are a lot of generated base clusterings which are connecting different clustering views in the clustering space. Thus, in the clustering space, they seems like a big meta-cluster. After filtering, we clean out these connecting base clusterings and the different clustering views are separated clearly. Thus, the iVAT diagram of the unfiltered base clusterings only contains one big dark block and two small blocks while the iVAT diagram of the filtered base clusterings contain multiple clear blocks. From the unfiltered base clusterings, the shape view is not discovered. It is because its meta-cluster is concealed in the big block. After filtering, the shape views is discovered and the quality of the color view is increased.

The MBM scores for clustering views generated from the unfiltered and filtered base clusterings are shown in Figure 38. In summary:

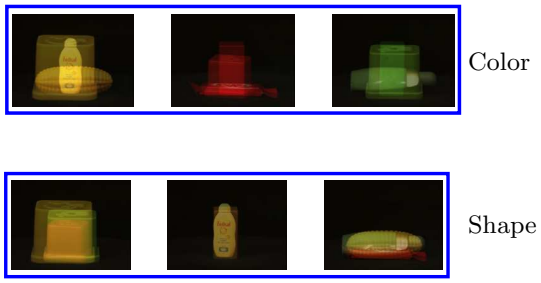
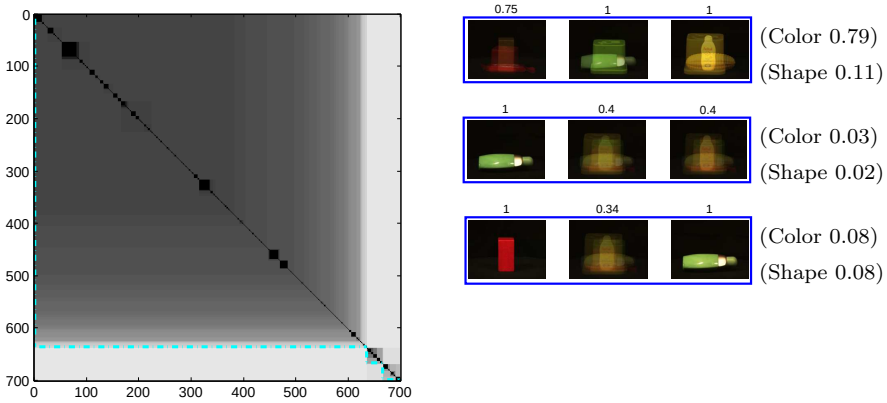


Fig. 35: Two ground truth clustering views on object dataset. The first row is the color view and the second row is the shape view.

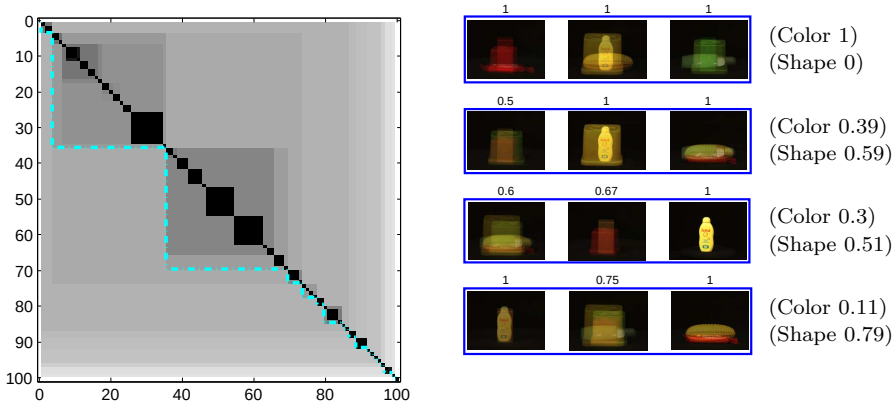


(a) The iVAT diagram of the unfiltered base clusterings on the object dataset.

(b) The top 3 clustering views got from the unfiltered base clusterings on the object dataset. The first row is the color view.

Fig. 36: The results on the unfiltered base clusterings on the object dataset.

1. We found out 3 clustering views from the unfiltered set of base clusterings and found out 9 clustering views from the filtered base clusterings with $L = 100, \beta = 0.95$.
2. The MBM scores for the 3 clustering views generated from the unfiltered base clusterings are invariant as only one color view is recovered and also the quality does not get better.
3. The returned top 4 clustering views from the filtered set of base clusterings recover and match well with the two ground truth views with $MBM(\mathcal{C}_4) = 0.9$.



(a) The iVAT diagram of the filtered base clusterings on the card dataset.

(b) The top 4 views got from the filtered base clustering on the object dataset. The first row is the color view and the fourth row is the shape view.

Fig. 37: The results for the filtered base clusterings on object dataset with $\beta = 0.3$.

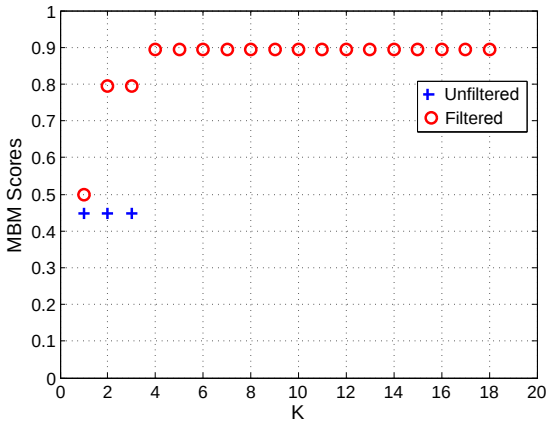


Fig. 38: MBM scores for clustering views generated from the unfiltered and filtered base clusterings on the object dataset.

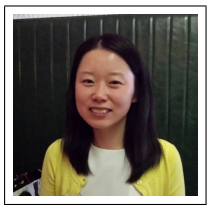
References

1. Azimi, J., Fern, X.: Adaptive cluster ensemble selection. In: IJCAI, vol. 9, pp. 992–997 (2009)
2. Bache, K., Lichman, M.: UCI machine learning repository (2013). URL `http://archive.ics.uci.edu/ml`
3. Bae, E., Bailey, J.: Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In: Data Mining, 2006. ICDM'06. Sixth International

Conference on, pp. 53–62. IEEE (2006)
4. Bailey, J.: Alternative clustering analysis: A review. In: C. Aggarwal, C. Reddy (eds.) Data Clustering: Algorithms and Applications. CRC Press (2013)
5. Caruana, R., Elhaway, M., Nguyen, N., Smith, C.: Meta Clustering. In: Proceedings of ICDM, pp. 107–118 (2006)
6. Cui, Y., Fern, X.Z., Dy, J.G.: Multi-view clustering via orthogonalization. In: Proceedings of ICDM, pp. 133–142 (2007)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886–893. IEEE (2005)
8. Dang, X.H., Bailey, J.: A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In: Proc. of KDD'10, pp. 573–582
9. Dang, X.H., Bailey, J.: Generating multiple alternative clusterings via globally optimal subspaces. Data Mining and Knowledge Discovery **28**(3), 569–592 (2014)
10. Dang, X.H., Bailey, J.: A framework to uncover multiple alternative clusterings. Machine Learning **98**(1-2), 7–30 (2015)
11. Davidson, I., Qi, Z.: Finding alternative clusterings using constraints. In: Proceedings of ICDM, pp. 773–778 (2008)
12. Faivishevsky, L., Goldberger, J.: Nonparametric information theoretic clustering algorithm. In: Proceedings of ICML, pp. 351–358 (2010)
13. Fern, X.Z., Lin, W.: Cluster ensemble selection. Statistical Analysis and Data Mining **1**(3), 128–141 (2008)
14. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD) **1**(1), 4 (2007)
15. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoretical Computer Science **38**, 293–306 (1985)
16. Gullo, F., Domeniconi, C., Tagarelli, A.: Metacluster-based projective clustering ensembles. Machine Learning **98**(1-2), 181–216 (2015)
17. Hadjitodorov, S.T., Kuncheva, L.I., Todorova, L.P.: Moderate diversity for better cluster ensembles. Information Fusion **7**(3), 264–275 (2006)
18. Havens, T.C., Bezdek, J.C.: An efficient formulation of the improved visual assessment of cluster tendency (ivat) algorithm. IEEE Transactions on Knowledge and Data Engineering **24**(5), 813–822 (2012)
19. Havens, T.C., Bezdek, J.C., Keller, J.M., Popescu, M.: Clustering in ordered dissimilarity data. Int. Journal of Int. Sys. **24**(5), 504–528 (2009)
20. Hossain, M.S., Ramakrishnan, N., Davidson, I., Watson, L.T.: How to "alternatize" a clustering algorithm. Data Mining and Knowledge Discovery **27**(2), 193–224 (2013)
21. Jain, A.K., Dubes, R.C.: Algorithms for clustering data (1988)
22. Jain, P., Meka, R., Dhillon, I.S.: Simultaneous unsupervised learning of disparate clusterings. Statistical Analysis and Data Mining: The ASA Data Science Journal **1**(3), 195–210 (2008)
23. Jaskowiak, P.A., Moulavi, D., Furtado, A.C., Campello, R.J., Zimek, A., Sander, J.: On strategies for building effective ensembles of relative clustering validity criteria. Knowledge and Information Systems **47**(2), 329–354 (2016)
24. Lei, Y., Vinh, N.X., Chan, J., Bailey, J.: Filta: Better view discovery from collections of clusterings via filtering. In: Machine Learning and Knowledge Discovery in Databases, pp. 145–160. Springer (2014)
25. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
26. Naldi, M.C., Carvalho, A., Campello, R.J.: Cluster ensemble selection based on relative validity indexes. Data Mining and Knowledge Discovery **27**(2), 259–289 (2013)
27. Nguyen, N., Caruana, R.: Consensus clusterings. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 607–612. IEEE (2007)
28. Nie, F., Wang, X., Huang, H.: Clustering and projected clustering with adaptive neighbors. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 977–986. ACM (2014)
29. Nie, F., Xu, D., Li, X.: Initialization independent clustering with actively self-training method. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(1), 17–27 (2012)
30. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1447–1454 (2006)

31. Niu, D., Dy, J.G., Jordan, M.I.: Iterative discovery of multiple alternativeclustering views. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(7), 1340–1353 (2014)
32. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1226–1238 (2005)
33. Phillips, J.M., Raman, P., Venkatasubramanian, S.: Generating a diverse set of high-quality clusterings. arXiv **1108.0017** (2011)
34. Pihur, V., Datta, S., Datta, S.: Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. Bioinformatics **23**(13), 1607–1615 (2007)
35. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
36. Sheng, W., Swift, S., Zhang, L., Liu, X.: A weighted sum validity function for clustering with a hybrid niching genetic algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **35**(6), 1156–1167 (2005)
37. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. The Journal of Mach. Learn. Res. **3**, 583–617 (2003)
38. Topchy, A., Jain, A.K., Punch, W.: Clustering ensembles: Models of consensus and weak partitions. IEEE Transactions on pattern analysis and machine intelligence **27**(12), 1866–1881 (2005)
39. Vinh, N.X., Epps, J.: minCEntropy: A novel information theoretic approach for the generation of alternative clusterings. In: Proc. of ICDM, pp. 521–530 (2010)
40. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of ICML, pp. 1073–1080. ACM (2009)
41. Wang, H., Shan, H., Banerjee, A.: Bayesian cluster ensembles. Statistical Analysis and Data Mining **4**(1), 54–70 (2011)
42. Wang, L., Nguyen, U.T., Bezdek, J.C., Leckie, C.A., Ramamohanarao, K.: iVAT and aVAT: enhanced visual analysis for cluster tendency assessment. In: Proceedings of PAKDD, pp. 16–27 (2010)
43. Zhang, Y., Li, T.: Extending consensus clustering to explore multiple clustering views. In: Proceedings of SDM, pp. 920–931 (2011)

Author Biographies



Yang Lei received a M.S. degree from Zhengzhou University, China, in 2012. She is currently a Ph.D. student in the Department of Computing and Information Systems, University of Melbourne, Australia. Her research interests include data mining and machine learning.

Dr. Vinh Nguyen is a Research Fellow at the Data Mining and Knowledge Discovery group at the University of Melbourne. His research focuses on the development and application of novel data mining and machine learning techniques across diverse problems in medical research, bioinformatics and computational biology, transportation and social networks. He has published in total 50+ research papers, many of which at top-tier venues in data mining and machine learning (SIGKDD, AAAI, ICML, IJCAI, SDM, ICDM, ECML, JMLR), attracting over 900 citations. He frequently serves as a program committee member and reviewer for many top-tier conferences and journals, including KDD, NIPS, ICDM, SDM, AAAI, JMLR. Currently, his research focuses on big data analytics and deep learning. Vinh is enthusiastic about applying machine learning and data science to solve real word problems, and is currently a Kaggle Master who ranks within the top 0.5% of all 50K+ Kagglers.

Dr. Jeffrey Chan is a Lecturer in the School of Science (Computer Science), at RMIT University, Australia. His research interests include graph and social network analysis, machine learning, social computing and dimension reduction. He holds a PhD in computer science from the University of Melbourne, Australia.

James Bailey is a Professor in the Department of Computing and Information Systems at the University of Melbourne. He has an extensive track record in databases and data mining and has been chief investigator on multiple ARC discovery grants and an Australian Research Council Future Fellow. He has been the recipient of six best paper awards and is an active member of the knowledge discovery community. He serves as an Associate Editor for the journals Knowledge and Information Systems and Social Network Analysis and Mining and has previously been as Associate Editor for IEEE Transactions on Knowledge and Data Engineering. He regularly serves as a Senior PC member for top conferences in data mining and was the co-general chair for ACM CIKM 2015 conference.