FILTA: Better View Discovery from Collections of Clusterings via Filtering

Yang Lei, Nguyen Xuan Vinh, Jeffrey Chan, and James Bailey

Department of Computing and Information Systems
University of Melbourne, Australia
yalei@student.unimelb.edu.au
{vinh.nguyen,jeffrey.chan,baileyj}@unimelb.edu.au

Abstract. Meta-clustering is a popular approach to find multiple clusterings in the datasest, which takes 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 this paper we propose a simple and effective filtering algorithm (FILTA) that can be flexibly used in conjunction with any meta-clustering method. Given a (raw) set of base clusterings, FILTA employs information theoretic criteria to remove those having poor quality or high redundancy. Then this filtered set of clusterings is highly suitable for further exploration, particularly the use of visualization for determining the dominant views in the dataset. We evaluate FILTA on both synthetic and real world datasets, and see how its use can enhance view discovery for complex scenarios.

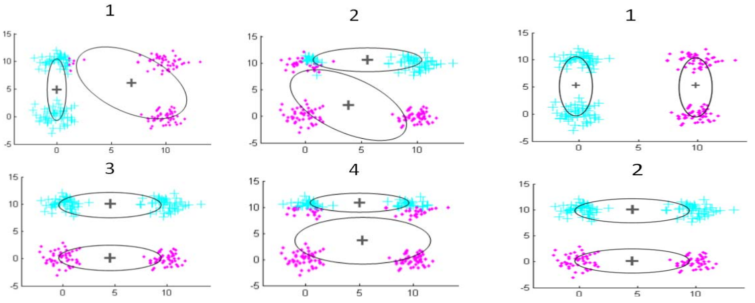
Keywords: Clustering, Meta-Clustering, Multiple Clusterings, Clustering Visualization.

1 Introduction

Clustering is one of the most important unsupervised techniques for discovering dataset structure. Many clustering methods focus on obtaining one single 'best' solution by optimizing a pre-defined criterion [11]. There are two limitations with this: firstly, data can be multi-faceted in nature. Particularly when the datasets are large and complex, there may be several useful clusterings that exist, not only 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* [2].

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 [3],[20]. Meta-clustering generates a large number of base clusterings using different procedures: running different clustering algorithms, running a specific algorithm several times with different initializations, or using random feature weighting in

T. Calders et al. (Eds.): ECML PKDD 2014, Part II, LNCS 8725, pp. 145–160, 2014.



(a) Four views generated from the raw base clusterings.

(b) Two views generated from the filtered base clusterings.

Fig. 1. Two sets of views found using as input a) raw set of *unfiltered* base clusterings, and b) set of *filtered* base clusterings. Colours indicate clusters.

the distance function. These base clusterings may then be meta-clustered into groups. Further, clusterings within the same group can be combined using consensus (ensemble) clustering to generate a consensus view of that group. This results in one or more distinctive clusterings (views) of the dataset, each offering a different perspective or explanation.

A major drawback and challenge with the use of meta-clustering is that its effectiveness is highly dependent on the quality and diversity of the generated base clusterings. Specifically, if the base clusterings have high similarity, then further processing may generate multiple similar views. If the base clusterings are of low quality, then naive visualization or analysis will produce low quality views. Users may be misled by these similar or poor quality views.

We illustrate this problem with an example in Figure 1a, where the dataset consists of four Gaussian clusters. We generate a set of raw base clusterings via k-means (with $k = 2$ clusters) and random feature weighting (these base clusterings are not shown in the figure). These base clusterings are then meta-clustered into groups, and for each group a consensus view is extracted via consensus clustering. The views generated on the raw base clusterings are presented in Figure 1a. Observe that among these four views, some are rather similar (view2, view3 and view4) and some have poor quality (view1, view2 and view4). This stimulates the following question, which is the basis for our paper - *Can we apply a filtering step to the base clusterings and thus avoid discovering poor quality or redundant views*? Figure 1b provides intuition about the benefits of filtering. It shows the views generated using a filtered set of base clusterings as input. These are more natural views of the dataset, being both of high quality and non-redundant.

In more detail, we propose filtered meta-clustering (FILTA), aiming at detecting multiple high quality and distinctive views by filtering and then 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

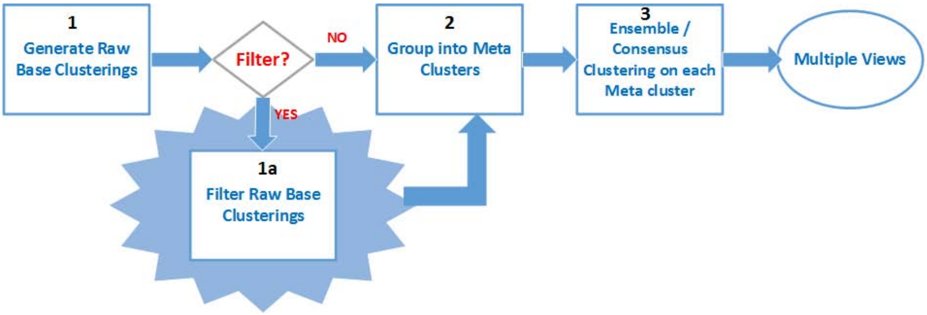

1
Generate Raw Base Clusterings
Filter?
NO
YES
1a
Filter Raw Base Clusterings
2
Group into Meta Clusters
3
Ensemble / Consensus Clustering on each Meta cluster
Multiple Views

Fig. 2. The meta-clustering framework with proposed filtering step highlighted

visual method to automatically determine the dominant meta clusters within the filtered base clusterings. Finally, we perform consensus clustering on each meta cluster to identify the views. Figure 2 shows the whole process. The novelty of our approach lies in the addition of a filtering step to the existing meta-clustering framework, which is highlighted as step 1a in Figure 2. Our focus is on investigating how to filter the given raw base clusterings to generate a set of better views, in terms of quality and diversity, compared to the unfiltered meta-clustering. We assume that we are given a set of base clusterings and the generation of appropriate base clusterings (step 1) is outside the scope of this paper and is left for future work. An important advantage of our method is that the filtering step is independent of the other steps in this framework and thus may be easily integrated with them.

Our contributions can be summarized as follows:

- We identify limitations with the current pipeline for meta-clustering. In particular, its reliance on the quality and diversity of a set of (raw) base clusterings for generating high quality and diverse views.
- We propose a novel *filtering* based meta-clustering approach for discovering multiple high quality and diverse views from a given set of base clusterings. Our filtering step can enhance any existing meta-clustering method.
- We propose a mutual information based filtering criterion which considers the quality and diversity of base clusterings simultaneously. We provide a parameter that allows users to flexibly control the balance between less number of views but of higher quality or more of them but of relatively lower quality.

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 views by generating and evaluating a large set of base clusterings. In work [3], 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 [20] proposed a method that extend [3] with consensus clustering in order to capture multiple views. Work in [14] proposed a sampling method for discovering a large set of good quality base clusterings. After that, the k-center [9] clustering method is used to select the k most dissimilar solutions as the views. 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. See [2] 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 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 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 [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 all the available solutions [8], 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 [8]. 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 2 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 the base clusterings. Alternative clustering employs objective functions to guide the search process, thus it may discover alternative views faster when compared to meta-clustering which employs a random clustering generation scheme (such as random initialization or random feature weighting). 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. 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.

3 FILTA: An Algorithm for Filtering Base Clusterings

Let us first introduce the notations used and a formal problem definition. 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 views be denoted by $\mathcal{V} = \{V_1, \ldots, V_R\}$, where a view V_i is a clustering on X, $V_i \in \mathcal{P}_{\mathcal{X}}$. Even though a view is just a clustering, we use the view nomenclature to distinguish between the initial base clusterings and the final, selected clusterings (the set of views) at the end of the meta-clustering process. 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 measure $Sim(C_i, C_j)$: $\mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}^+$. Our problem can be formalized as follows.

Problem Definition 1 *Given a set of raw base clusterings $\mathcal{C} = \{C_1, \ldots, C_l\}$, we seek a set of 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 $\mathcal{C}$. Next we discuss the quality and diversity criteria for clusterings.

3.1 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 [5]. For measuring similarity between clusterings, mutual information can detect any kind of linear or non-linear relationship between random variables [17]. More specifically, the quality of a clustering C is measured by the amount of shared information with the data X, i.e., $I(X; C)$. Intuitively, the more information that is shared, the better that a clustering models the data. The mutual information between two clusterings $I(C_i; C_j)$ quantifies their similarity. Thus, the less mutual information shared, the more dissimilar they are. 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(·) is the Shannon entropy function). The diversity can be optimized by minimizing the average similarity between clusterings, as:

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

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 $H(X|C)$ [7], due to its desirable properties of efficiently estimating density functions in high dimensional data and being parameterless. The mutual information between two clusterings C_i and C_j is computed directly from their contingency table:

$$I(C_i;C_j) = \sum_{u \in C_i} \sum_{v \in C_j} p(u,v) \log \frac{p(u,v)}{p(u)p(v)} \tag{2}$$

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.

3.2 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 [13] which maximizes feature relevancy while minimizing feature redundancy, we propose a clustering selection criterion which combines the quality and diversity of clusterings:

$$\min_{\mathcal{C}' \subset \mathcal{C}, |\mathcal{C}'| = L} \left\{ \frac{1}{|\mathcal{C}'|} \sum_{C_i \in \mathcal{C}'} H(X|C_i) + \frac{\beta \beta_0}{|\mathcal{C}'|^2} \sum_{C_i, C_j \in \mathcal{C}', i \neq j} I(C_i;C_j) \right\} \tag{3}$$

where L is a user defined parameter specifying the number of base clusterings $\mathcal{C}'$ to be selected, and $\beta \in [0, \infty)$ is a trade-off parameter that balances the quality and diversity during selection. To make sure the second term is on the same scale as the first term, we set $\beta_0 = \max H(X|C_i) / \max I(C_i;C_j)$. 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 (3).

A simple incremental search strategy can be used to select a good subset $\mathcal{C}'$ for the criterion (3) 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:

$$\arg\min_{C_i \in \mathcal{C} \setminus \mathcal{C}'} \left\{ H(X|C_i) + \frac{\beta \beta_0}{|\mathcal{C}'|} \sum_{C_j \in \mathcal{C}'} I(C_i;C_j) \right\} \tag{4}$$

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. The overall computational complexity of the filtering step costs $O(|\mathcal{C}| \cdot n^2 d)$, where n is the number of data observations and d is the number of data features.

4 Discovering the Clustering Views

We have obtained a filtered set of base clusterings after performing the filtering process. Next we group them into clusters at the meta level (step 2 in Figure 2) and then perform ensemble clustering on each meta cluster for view generation (step 3). We first explain the measure used to compute the similarity between the base clusterings, then explain a visualization technique called VAT for determining the potential number of meta clusters. We then introduce a method that combines with VAT 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 for pairwise clustering comparison. Several measures of clustering similarity have been proposed in the literature [11]. Here, we utilize the Adjusted Mutual Information(AMI) [18], which is an adjusted-for-chance version of the normalized mutual information [16]. The AMI between two clusterings C_i and C_j is defined as:

$$AMI(C_i;C_j) = \frac{I(C_i;C_j) - E\{I(C_i;C_j)\}}{\max\{H(C_i),H(C_j)\} - E\{I(C_i;C_j)\}} \tag{5}$$

where the $E\{\cdot\}$ is the expected value of mutual information $I(C_i;C_j)$, and $H(C_i)$ 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 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 [19] 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 our work, each clustering can be taken as a data object, and we utilize the VAT method to visualize the number of potential meta clusters.

For grouping the set of clusterings, existing research uses hierarchical clustering [3],[20]. Our FILTA algorithm is not restricted to any particular grouping method. Here, we employ an automatic clustering method-CLODD which automatically extracts the number of clusters and produces a hard partition of the data objects based on a reordered dissimilarity matrix [10]. We obtain the reordered dissimilarity matrix by applying the VAT method to the dissimilarity matrix of clusterings. 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.

Discovering the Views via Ensemble Clustering: In this final step, we use the MCLA ensemble clustering algorithm [16] to find a consensus view for each meta cluster. At the end of this step, we have a set of high quality and diverse views of the data.

5 Experimental Results

In this section, we use a synthetic and two real world datasets to compare the performance of our FILTA method against the existing meta-clustering methods, i.e., we compare the views generated from the filtered base clusterings against the views discovered from the raw base clusterings.

Experimental Setup: We generate 400 base clusterings for each dataset. Then FILTA and the proposed steps in Section 4 are applied. The base clusterings are generated using a combination of the following six clustering methods, some of which have been used previously in other meta-clustering algorithms:

- k-means with random initializations.
- random feature weighting method where feature weights are drawn from the zipf distribution[3].
- 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 [15] 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.
- an alternative clustering method, minCEntropy [17], with different reference clusterings generated by one of the above methods.

5.1 Evaluation of the Resulting Views

In our experiments, we use two measures for evaluating the discovered views. The Dunn Index is a popular internal clustering quality measure [6] and is defined as: $DI(C) = \frac{\min_{i \neq j}\{\delta(c_i,c_j)\}}{\max_{1 \leq w \leq k}\{\Delta(c_w)\}}$, where δ is the cluster to cluster distance and Δ is the cluster diameter. A larger DI is better. When we seek to compare against the ground truth labels, we use the adjusted mutual information (AMI).

Inspired by Mean Average Precision(MAP) [12], a popular measure for evaluating ranked retrieval of documents in information retrieval, we propose a mean best matching (MBM) score to test: i) (diversity) how many ground truth labels can be recovered by the top k views? and ii) (quality) how well do the top k views match the multiple sets of ground truth labels? Here, we select the top k views according to their quality (measured by the DI) and then we assess the matching between these views and the ground truth labels using AMI. In more detail, given multiple ground truth views $\mathcal{G} = \{G_1, \ldots, G_H\}$ and a set of ranked views $\mathcal{V}_r = \{V_{r_1}, \ldots, V_{r_m}\}$, the mean best matching score for the top k views $\mathcal{V}_{r_k} = \{V_{r_1}, \ldots, V_{r_k}\}$, where $k \leq m$, is defined as:

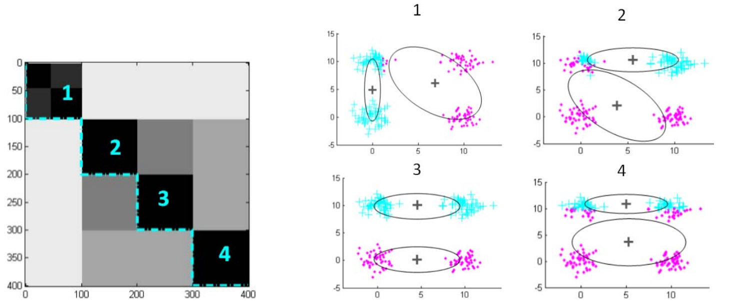
$$MBM(\mathcal{V}_{r_k}) = \sum_{i=1}^{H} \max_{V_j \in \mathcal{V}_{r_k}} AMI(G_i, V_j)/H \tag{6}$$

5.2 Synthetic Dataset

In this section, we use a synthetic dataset to test whether our FILTA method is able to discover high quality and diverse views by filtering out poor quality and similar base clusterings. Our synthetic dataset consists of four Gaussian clusters. Each of the generated 400 raw base clusterings consists of two clusters. There are two high quality and dissimilar views within these base clusterings and we aim to recover these.

Using the unfiltered set of base clusterings, the CLODD method produced four meta clusters, highlighted by the green dashed line surrounding the blocks in Figure 3a. After performing the ensemble clustering on each meta cluster, four views are generated (see Figure 3b), with their numbers corresponding to the numbered blocks in the VAT diagram. We can see that views 2, 3 and 4 are similar and redundant, while views 1, 2 and 4 are of poor quality (the consensus clusters are spread out) and only view 3 is of good quality. Note that views 1 and 3 correspond to the two ground truth views included in the raw base clusterings. However, view 1 is of low quality, due to poor quality base clusterings included in its meta cluster. This experiment demonstrates that the meta-clustering methods may generate poor quality and similar views since it uses all the base clusterings, whether they are of high quality or not.

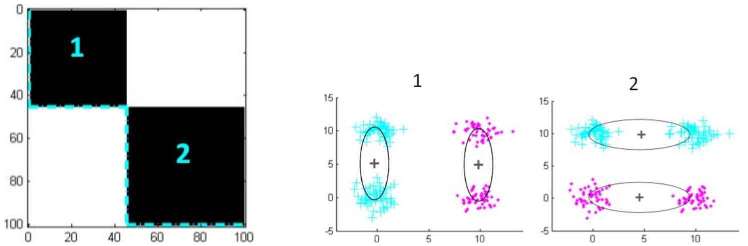
Next, we apply our FILTA algorithm on the same set of 400 base clusterings. We filter out 300 of the low quality and similar base clusterings setting $L = 100$ and $\beta = 0.1$ and the results are shown in Figure 4. The corresponding VAT diagram is presented in Figure 4a. Observe that there are two clearly separated blocks, indicating that there are two groups of clusterings that exist in the filtered base clusterings. The views generated based on the discovered two groups are presented in Figure 4b. We can see that these two views are of high quality, dissimilar and correspond to the two ground truth views. In addition, we can



(a) VAT diagram of the 400 base clusterings.

(b) Four discovered views, numbered according to the diagonal blocks of Figure 3a.

Fig. 3. Four views discovered from the 400 raw base clusterings of the synthetic dataset. Each numbered block in Figure 3a represents a view.



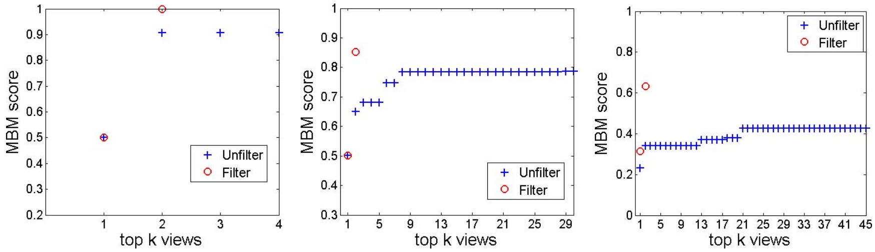
(a) VAT diagram of the 100 filtered base clusterings.

(b) Two discovered views, each corresponding to a block in Figure 4a.

Fig. 4. Two views discovered from the 100 filtered base clusterings of the synthetic dataset. Each numbered block in Figure 4a represents a view.

also observe that view 1 generated by FILTA has better quality compared to the view 1 generated by the unfiltered method (Figure 3b). It is because our filtering method filtered out the poor quality base clusterings in this group.

For a quantitative comparison of both approaches, we plot the MBM scores for the top 4 views in Figure 5a. Recall that FILTA only produced 2 views, hence it only has 2 scores in Figure 5a. Observe that for the top 1 view, both methods achieve the same MBM score, which means that the first view for each of these methods is of the same quality. For the top 2 views, the MBM score for the unfiltered method is lower than the FILTA method, because the view generated by the unfiltered method has a relative lower quality than the one generated by FILTA method. As we can see, with the increasing of value k, the MBM score for the unfiltered method does not change. It is because the third and fourth views of the unfiltered method do not perform better than the top 2 views meaning that they are either similar views to the others (redundant) or of poor quality.



(a) Synthetic dataset. **(b)** CMUFace dataset. **(c)** Isolet dataset.

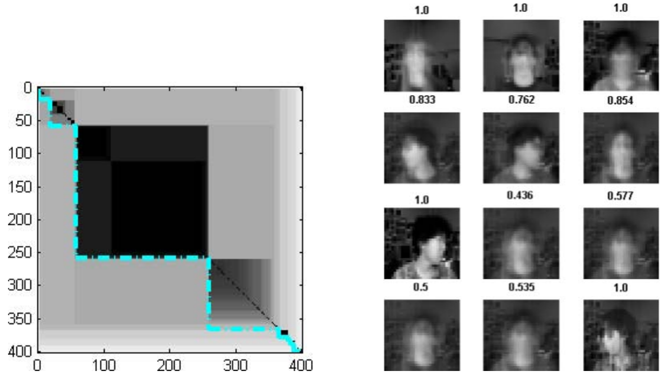
Fig. 5. The mean best matching (MBM) scores for the top k views of three datasets. One set of views is generated from the raw base clusterings (results represented by blue crosses), while the other set is from filtered ones (red circles).

5.3 CMUFace Dataset

We next show how FILTA performs for two real datasets. The CMUFace dataset from the UCI Machine Learning Repository [1] is a commonly used dataset for the discovery of alternative clusterings [4]. It contains 624 32×30 images of 20 persons, along with different features of these persons, e.g., pose (straight, left, right, up). Two dominant views exist in this dataset - identity and pose. In our experiment, we randomly select the images of three people and have 93 images in total. Again we generated 400 base clusterings and FILTA selected $L = 100$ with $\beta = 0.03$.

The results generated by the unfiltered method are shown in Figure 6. From the VAT diagram in Figure 6a, we can observe that there are a larger number of diagonal blocks. The CLODD algorithm produces 30 meta clusters and hence we have 30 views overall. Due to the limitation of space, we just show the top 4 views as measured by DI in Figure 6b. Each row in the figure is a view of three clusters, and each cluster of images is illustrated by its mean image. The number above each image is the purity score, which is the percentage of images, of a cluster, with the majority ground truth label (this can be labels from the identity or pose views). Higher purities are desirable. Consider Figure 6b. The view displayed in the first row corresponds to the person ground truth view, and the view in the second row corresponds to the pose one. However, the third and fourth row views are a combination of the other two. Their clusters mix poses and identities. From this experiment, we can see that the existing unfiltered meta-clustering methods can generate many poor quality and redundant views.

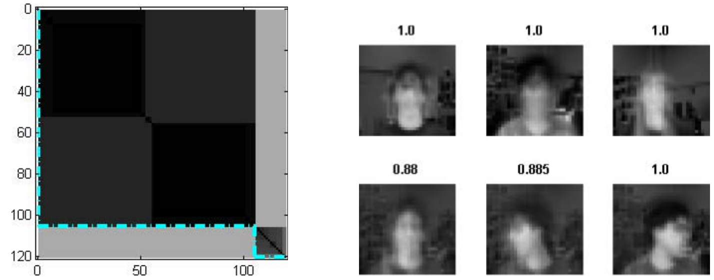
The results generated by our FILTA method are shown in Figure 7. As we can observe, the VAT diagram (Figure 7a) is less fuzzy compared with the one generated from the raw base clusterings (Figure 6a), has higher purity and includes two relatively well separated blocks. Again our filtering method has filtered out the poor quality and redundant base clusterings. Two views are generated according to the discovered two groups shown in the VAT diagram, and are shown in Figure 7b. They are the desired person and pose views. Compared with the



(a) VAT diagram of the raw base clusterings.

(b) Four views with highest DI.

Fig. 6. Views generated from the 400 raw base clusterings of the CMUFace data. The number above each image is its purity score.



(a) VAT diagram of the filtered base clusterings.

(b) Two views generated from the filtered base clusterings.

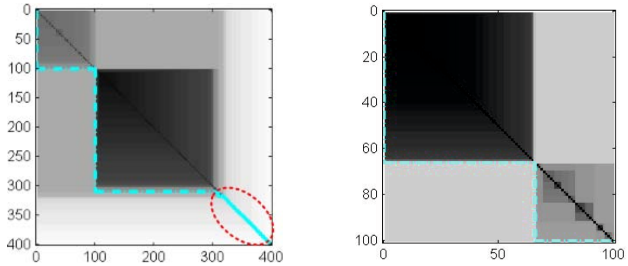
Fig. 7. Views generated from the 100 filtered base clusterings of the CMUFace data. The number above each image in Figure 7b is its purity score.

pose view generated by the unfiltered method (Figure 6b), we get better quality in terms of the purity score shown above the image.

The MBM scores for these two sets of views are shown in Figure 5b. As we can see, the best MBM score for the unfiltered method is reached at the 8th view, implying that noisy results are present in the top 7 views. This result further demonstrates the influence of the quality and diversity of the base clusterings on the performance of the unfiltered meta-clustering methods.

5.4 Isolet Dataset

The isolet dataset from UCI machine learning repository [1] contains 7797 records with 617 features, which come from 150 subjects speaking the name of each letter of the alphabet twice. There are two views (speaker and letters) in this dataset.



(a) VAT diagram of the 400 raw base clusterings.

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

Fig. 8. VAT diagrams for the raw and filtered sets of base clusterings of the Isolet dataset

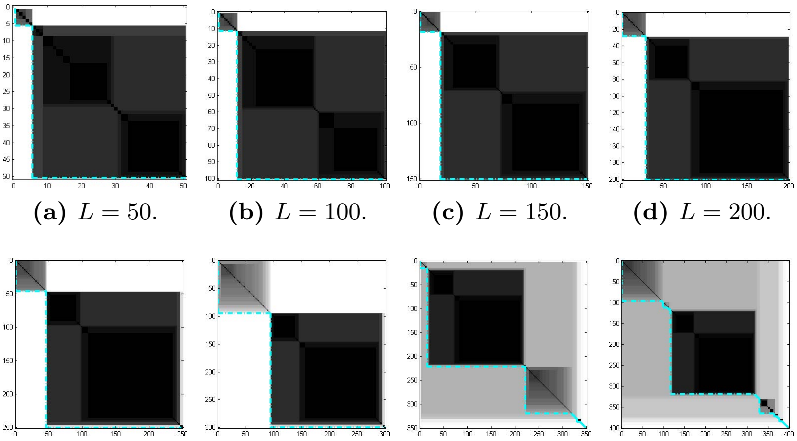
In our experiment, we randomly selected 10 persons along with 10 letters, resulting in a 200 records dataset. We generate 400 base clusterings that contains the speaker and letter views, and select 100 base clusterings using FILTA ($\beta = 0.5$).

The results are shown in Figure 8. From the VAT diagram on the raw base clusterings (Figure 8a), we can observe that there are many small, dissimilar blocks in the right bottom corner of the VAT diagram, highlighted by the red dashed circle (dissimilarity indicated by the light shading of the area between the blocks). Each of them is taken as a view which results in 45 views overall. After applying our filtering method on the raw base clusterings, we obtain the VAT diagram in Figure 8b. As we can see, there are two explicit big blocks without those dissimilar individual meta clusters, which have been filtered out due to their poor quality.

The MBM scores for these two sets of views are shown in Figure 5c. It can be observed that the top 1 view generated by our FILTA method has higher quality than the one generated by the unfiltered method. In addition, the two views of FILTA capture the two ground truth views well. In contrast, the existing unfiltered method generated almost 45 views and the quality of the best matching views for the two ground truth views among the 45 views are not comparable with FILTA's. This result further shows that the input base clusterings, including low quality and redundant solutions, will lead to similar and poor quality views.

5.5 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 CMUFace dataset as the example to show the impact of L. In Figure 9, we show the VAT diagram constructed for $L = 50$ to 400 (filtered) base clusterings (recall that there are 400 raw base clusterings). We see that the VAT diagrams are mostly stable from $L = 50$ to 300, meaning that FILTA is quite robust to noise and relatively insensitive to the choice of L. From our experiments we found $L = 25\% \times l$ (l is the number of raw base clusterings) works well.



(a) $L = 50$. **(b)** $L = 100$. **(c)** $L = 150$. **(d)** $L = 200$.

(e) $L = 250$. **(f)** $L = 300$. **(g)** $L = 350$. **(h)** $L = 400$.

Fig. 9. VAT diagrams for different number of filtered base clusterings with $\beta = 0.03$ on the CMUFace data

5.6 Impact of the Regularization Parameter

The regularization parameter $\beta \in [0, \infty)$ balances the quality and diversity during the clustering filtering procedure. We have found that $\beta \in [0, 1]$ works fairly well. Essentially within this range, we place more emphasis on clustering quality. For example, when $\beta = 0.5$, it means we treat quality as twice as important as diversity. When $\beta \to 1$, the filtering process places equal emphasis on diversity, which generally increases the number of potential views but at the risk of including more poor quality solutions. In contrast, when $\beta \to 0$, the filtering procedure focuses on the quality, which will result in high quality views but some relevant views may be filtered out. Thus, users can tune this parameter according to their specific needs for view detection. Given that we usually do not have the cluster labels, the VAT 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 VAT reordered matrix and the consensus views that emerge. We demonstrate the effect of β on the CMUFace dataset. Figure 10 shows how the MBM score changes as we vary β. As it can be seen, a $\beta = 0.03$ to 0.05 gives the best matching scores. To further confirm these are effective β values for this dataset, we illustrate a number of VAT diagrams (Figures 11) constructed from different β values and $L = 100$. The diagrams show that $\beta = 0.03$ (Figure 11b) discovers two relatively sharp dark blocks which are turned out to correspond to the two true views. Also, we can see that as β increases, the VAT diagram becomes more fuzzy, which means that the selected base clusterings are more diverse but their quality is decreasing. In this respect, our proposed framework is a useful tool to assist the discovery of novel views from the data.

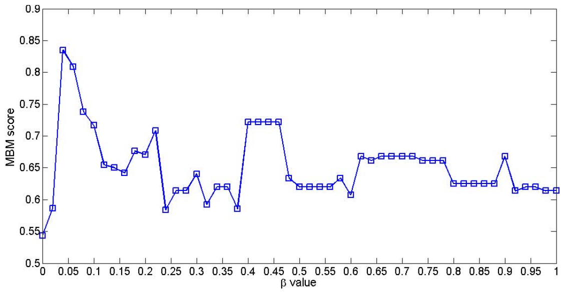
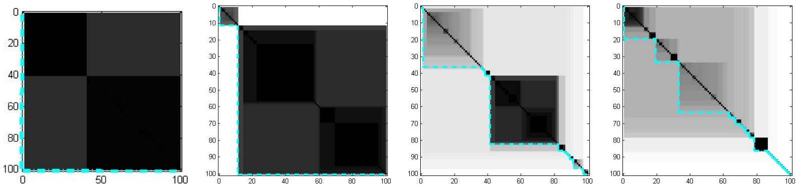


Fig. 10. The mean best matching scores(MBM) with different β on 100 filtered clusterings generated from CMUFace dataset



(a) $\beta = 0$ **(b)** $\beta = 0.03$ **(c)** $\beta = 0.1$ **(d)** $\beta = 1$

Fig. 11. VAT diagrams generated from 100 filtered base clusterings and different β values, for the CMUFace data

6 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 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. We believe FILTA is a simple and useful tool to incorporate within the area of multiple clustering exploration and analysis.

References

1. Bache, K., Lichman, M.: UCI machine learning repository (2013)
2. Bailey, J.: Alternative clustering analysis: A review. In: Aggarwal, C., Reddy, C. (eds.) Data Clustering: Algorithms and Applications. CRC Press (2013)
3. Caruana, R., Elhaway, M., Nguyen, N., Smith, C.: Meta Clustering. In: Proceedings of ICDM, pp. 107–118 (2006)
4. Cui, Y., Fern, X.Z., Dy, J.G.: Multi-view clustering via orthogonalization. In: Proceedings of ICDM, pp. 133–142 (2007)
5. Dang, X.H., Bailey, J.: A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In: Proc. of KDD, pp. 573–582 (2010)
6. Davidson, I., Qi, Z.: Finding alternative clusterings using constraints. In: Proceedings of ICDM, pp. 773–778 (2008)
7. Faivishevsky, L., Goldberger, J.: Nonparametric information theoretic clustering algorithm. In: Proceedings of ICML, pp. 351–358 (2010)
8. Fern, X.Z., Lin, W.: Cluster ensemble selection. Statistical Analysis and Data Mining 1(3), 128–141 (2008)
9. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoretical Computer Science 38, 293–306 (1985)
10. 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)
11. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice-Hall, Inc. (1988)
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press, Cambridge (2008)
13. 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)
14. Phillips, J.M., Raman, P., Venkatasubramanian, S.: Generating a diverse set of high-quality clusterings. arXiv, 1108.0017 (2011)
15. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
16. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. The Journal of Mach. Learn. Res. 3, 583–617 (2003)
17. 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)
18. 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 (2009)
19. Wang, L., Nguyen, U.T.V., Bezdek, J.C., Leckie, C.A., Ramamohanarao, K.: iVAT and aVAT: Enhanced visual analysis for cluster tendency assessment. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010, Part I. LNCS, vol. 6118, pp. 16–27. Springer, Heidelberg (2010)
20. Zhang, Y., Li, T.: Extending consensus clustering to explore multiple clustering views. In: Proceedings of SDM, pp. 920–931 (2011)