A multiplex-network based approach for clustering ensemble selection

Parisa Rastin, Rushed Kanawati
LIPN CNRS UMR 7030, UP13-SPC
99 Av. J.B. Clément 9430 Villetaneuse, France
Email: surname.name@lipn.univ-paris13.fr

Abstract—Performance of cluster ensemble approaches is now known to be tightly related to both quality and diversity of input base clusterings. Cluster ensemble selection (CES) refers to the process of filtering the raw set of base clusterings in order to select a subset of high quality and diverse clusterings. Most of existing CES approaches apply one index for measuring the quality and another for evaluating the diversity of clusterings. Moreover the number of clusterings to select is usually given as an input to the CES function. In this work we propose a new CES approach that allow taking into account an ensemble of quality and diversity indexes. In addition, the proposed approach computes automatically the number of clusterings to return. The basic idea is to define a multiplex network over the provided set of base clusterings. Each slice in the multiplex network is obtained by defining a proximity-graph over the set of base clusterings using a given clustering dissimilarity index. A community detection algorithm is applied to the obtained multiplex network. We then rank clusterings in each community applying an ensemble-ranking approach using different (internal) clustering quality indexes. From each community we select the base clustering ranked at the top. First experiments on benchmark datasets shows the effectiveness of the proposed CES approach.

Index Terms—Ensemble clustering, Clustering ensemble selection, Multiplex network, Community detection.

I. INTRODUCTION

Ensemble clustering is a meta-learning approach that consists in combining multiple clusterings defined on the same set of data items [1]. Let \( \Pi = \{ \pi_1, \ldots, \pi_n \} \) a set of \( n \) clusterings defined on the same set of items. We call \( \Pi \) the set of base clusterings. An ensemble clustering (EC) approach is defined as follows:

\[
EC(\Pi) = \arg \min_{\pi^*} \sum_{i \in [1,n]} d(\pi^*, \pi_i)
\]

Where \( d() \) a clustering dissimilarity function. In other words, an EC approach computes a consensus clustering that minimizes disagreement with each of the input base clusterings. Notice that an EC approach does not require accessing features of clustered data neither those of clustering algorithms used to generate each of input base clusterings. This makes EC approaches suitable to combine clusterings obtained by different algorithms, or by different configurations of the same algorithm. It can also be used to combine clusterings obtained by using different features of the data items (i.e. multi-view clusterings). It has been also used to overcome stability issues in data clustering [2].

More recently, EC approaches have been also used in the hot field of community detection in complex networks. Actually, real complex networks exhibit a community structure where the network can be divided into sub-graphs (i.e. communities) that are highly dense but loosely linked to other sub-graph in the network [3]. EC approaches have been applied for different tasks including: computing communities cores [4], computing dynamic communities [5], multi-objective local communities identification [6], community detection in multiplex networks [7], and large-scale graph coarsening [8], [9].

Different consensus clustering functions have been proposed in the literature. Existing functions can be roughly classified into two classes: evidence accumulation based functions [10] and graph-based functions [1]. One widely applied method is the CSPPA approach. The approach is based on constructing a consensus graph out of the set of partitions to be combined [11], [1]. The consensus graph \( G_{cons} \) is defined over the same set of clustered data items. Two nodes \( v_i, v_j \in V \) are linked in \( G_{cons} \) if there is at least one base partition where both items \( i,j \) are in a same cluster. Each link \( (v_i, v_j) \) is weighted by the frequency of instances that nodes \( v_i, v_j \) are placed in the same cluster. Links in the obtained consensus graph whose weights (frequency) are under a given threshold \( \alpha \in [0,1] \) are pruned yielding decomposing the graph in a set of connected components. These connected components represent the consensus clustering.

Recently, different works have showed that the quality of the output of an EC approach is tightly related to both the quality of each partition in the base clustering set and diversity of these clusterings. Cluster ensemble selection (CES) approaches have been proposed in order to compute a subset of the base clusterings set that maximize both the quality and the diversity.

The diversity of clusterings can be estimated by applying different external cluster evaluation indexes or cluster dissimilarity indexes such as: rand index and the adjusted rand index (ARI) [12], Normalized mutual information (NMI) and Information variation indexes [13]. If clustered data are structured in the form of a graph, specific versions of these external evaluation indexes can also be applied [14].

The quality of a clustering can be evaluated by different internal evaluation indexes. Examples are: Silhouette index, Calinski-Harabasz index, Davis-Bouldin index and Dunn index [15]. If raw clustered data are graphs, then community evaluation metrics can then be applied. Examples are the
modularity [16], or the different local modularities functions [17], [6].

Almost all CES approaches require the number of base clusterings to select as an input [18]. Some are based on selecting high quality base clusterings. Some compute a trade-off between quality and diversity [19]. However, all existing approaches apply one index for measuring the quality and another for evaluating the diversity of clusterings. In this work we investigate the use of an ensemble of quality and diversity indexes for CES. In addition the devised approach computes automatically the number of base clusterings to be selected. This will be explained in more details in next section.

II. MULTIPEX NETWORK BASED CES

Algorithm 1 sketches the outlines of the proposed approach. The basic idea is to define a multiplex network over the provided set of base clusterings. A multiplex network is multi-slice network where each slice contains the same set of nodes but different kinds of links [20]. In our case, each slice of the defined multiplex is modeled by a proximity graph constructed by applying a given clustering dissimilarity index (ex. NMI, ARI, VI). Different types of proximity graphs can be used. In this work, we first explore using relative neighbourhood graphs (RNG) [21]. Though the complexity of RNG graph construction is relatively high, the resulted graph is proved to be connected and sparse. Recall also that the graph is defined over the set of base clusterings which cardinality is usually low.

Algorithm 1 Graph-based cluster ensemble selection algorithm

[Algorithm code]

A community detection algorithm is applied to the obtained multiplex network. Recall that a community is defined as a dense sub-graph that is loosely connected to other communities in the network. Different approaches for community detection in multiplex networks can be applied. A survey on such algorithms is provided in [22]. In this work we apply a seed-centric approach proposed in [20]. Since two nodes (clusterings) are linked if they very similar, a community in the multiplex network delimits a number of base clusterings that are similar among them and diverse in regard to other clusterings belonging to other communities. We can then stress the diversity of clusterings to return by selecting one clustering from each detected community. We rank clusterings in each community applying an ensemble-ranking approach using different (internal) clustering quality indexes. From each community we select the base clustering ranked at the top. First experiments on benchmark datasets shows the effectiveness of the proposed CES approach. This leads to select high quality but diverse base clusterings.

III. EXPERIMENTS

As a first evaluation of the proposed CES approach, we have conducted the following primary experiment. We have selected a set of benchmark networks frequently used in works dealing community detection in complex networks and for which we have a ground-truth decomposition into communities. These networks are the following: the Zachary Karate club network, the US politics books network and Dolphins network [6]. To each network, we apply the label propagation community detection algorithm 100 times [23]. This algorithm is known to be quick but highly instable. We then obtain a set of 100 different clusterings that compose our raw base clusterings set. We then compared the results of applying a CSP for clustering approaches directly to the raw base clusterings set to those obtained by applying the same ensemble clustering algorithm to the subset obtained after applying our CES approach. For the CES algorithm, we used the modularity, and the conductance as a clustering quality indexes. NMI, ARI and VI are used to measure clustering dissimilarity (diversity). The muxLicod algorithm [20] is applied in order to compute communities in the obtained multiplex network. A simple Borda rank aggregation method is applied in order to select the top quality clustering from each detected community. The results are evaluated in function similarity of obtained clustering to the ground-truth clustering using again the NMI and ARI indexes. The modularity (Q) is also used to evaluate the overall quality of obtained results. As shown in next table, for all three networks, the CES approach does enhance the obtained results. These first results are encouraging. But the work is still in its early stages. Experiments on larger datasets and using different quality and diversity indexes are scheduled. The effect of the choice of the multiplex community detection algorithm should also be studied. Another factor to analyse is the enhancement in using an ensemble of indexes rather than using single quality/diversity index should also be done.

IV. CONCLUSION

In this work, we have proposed a new approach for enhancing the output of ensemble clustering by applying an original ensemble selection process. The approach consists in applying a community detection algorithm to a multiplex graph defined over the set of base clustering to filter. First results show that the overall quality of obtained clustering is enhanced when applying ensemble selection process. Experiments on
large-scale datasets are planned in order to confirm these first but promising results. Comparisons with other ensemble selection approaches based on implicit quality estimation are also scheduled.

REFERENCES


a) Parisa Rastin: is currently a master-II student at university Paris 13. She is working on ensemble clustering and clustering ensemble selection approaches as part of her master degree internship.

b) Rushed Kanawati: is associate professor at LIPN, university Paris 13 since 2000. His current research interests are mainly in the field of complex network analysis, machine learning approaches and recommender systems. His current research interests cover, community detection algorithms, link prediction, multiplex network analysis and ensemble learning approaches.