Ensemble Selection for Community Detection in Complex Networks

Rushed Kanawati(✉)

LIPN, CNRS UMR 7030, University Paris 13, Sorbonne Paris Cité,
Villetaneuse, France
rushed.kanawati@lipn.univ-paris13.fr
http://www-lipn.univ-paris13.fr/kanawati

Abstract. Ensemble clustering approaches have been recently applied, in a variety of ways, in order to enhance the quality and/or the execution time of community detection tasks. The quality gain that can be obtained from applying ensemble approaches is known to be tightly linked to both quality and diversity of the applied clusterings. However, most of existing work simply ignore this important issue of ensemble selection. In this paper we intend to fill this gap. We propose a graph-based ensemble selection approach that allow to take into account both criteria of quality and diversity. Different quality measures are also considered: cluster-oriented quality and network-oriented quality functions. Experiments on real network datasets show the validity of our approach.

Keywords: Community detection · Complex networks · Ensemble clustering · Ensemble selection

1 Introduction

Complex networks are frequently used for modeling interactions in real-world systems in diverse areas, such as sociology, biology, information spreading and exchanging and many other different areas. One key topological feature of real-world complex networks is that nodes are arranged in tightly knit groups that are loosely connected one to each other. Such groups are called *communities*. Nodes composing a community are generally admitted to share common proprieties and/or be involved in a same function and/or having a same role. Hence, unfolding the community structure of a network could give us much insights about the overall structure a complex network. Comprehensive review of the state of the art can be found in [6,24]. Different algorithms have different execution times and yield results of various quality.

The large-size of today available networks makes most of existing algorithms hard to apply. In addition, most of existing low time complexity algorithms show generally low robustness. Different executions of the same algorithm on the same network may leads to detecting highly different partitions of the network. This is for instance the case of the *Louvain* approach [2] which is sensitive to the order

G. Meiselwitz (Ed.): SCSM 2015, LNCS 9182, pp. 138–147, 2015.
DOI: 10.1007/978-3-319-20367-6_15

in which nodes of the network are parsed. Another very known exemple is the high speed label propagation algorithm the exhibits, in its original version [18], a very high instability.

Ensemble clustering approaches have been proposed as a mean for both graph coarsening and graph clustering enhancing. Graph coarsening refers to the process of reducing the scale of a graph by replacing a group of cohesive nodes in the graph by a single node [22]. High quality community detection algorithms, with higher computational complexity, can then be applied on the reduced graph. Results are then expanded to the initial graph. Ensemble clustering can directly be applied in order to merge different clustering obtained by applying different algorithms or by applying an unstable algorithm several times [21]. However, the quality gain that can be obtained from applying ensemble approaches is known to be tightly linked to both quality and diversity of the applied clusterings [1,4]. Most of existing work simply ignore this important issue of ensemble selection. In this paper we intend to fill this gap.

The remainder of this paper is organized as follows. Next in Sect. 2, we first define the problem of ensemble clustering, discuss main approaches for consensus clustering computation and show applications in the field of community detection in complex networks. In Sect. 3 we define the problem of ensemble selection and quickly review main ensemble selection approaches. The proposed graph-based ensemble selection algorithm is presented in Sect. 3.2. Experiments and results are reported and commented in Sect. 4. Finally we conclude in Sect. 5.

2 Applying Ensemble Clustering to Community Detection

2.1 Ensemble Clustering Approaches

Let $G =< V, E >$ be a undirected simple graph where V is the set of nodes and E is the set of edges. Let π_i be a partition of the set V. We have by definition $\pi_i = \{\pi_i^1, \ldots, \pi_i^l\}$ where $\pi_i^j \subseteq V$, and $\bigcup_j \pi_i^j = V$ and $\forall j, k \in [1, l] \pi_i^j \cap \pi_i^k = \emptyset$.

We consider a set of a different partitions $\mathcal{P} = \{\pi_1, \ldots, \pi_n\}$ defined over the same set V. The goal of an ensemble clustering function is to compute a consensus clustering π_* that minimize the number of disagreements with each base partition π_i. In a formal way we have:

$$\pi_* = \arg\min_{\pi_i \in \mathcal{P}} dist(\pi_*, \pi_i) \tag{1}$$

Where $dist()$ is a distance function measuring disagreement between two partitions. Some exemples of such distance functions are given in Sect. 3.

Different consensus clustering functions have been proposed in the literature. Existing functions can be roughly classified into two classes: *evidence accumulation based functions* [7] and *graph-based functions* [23]. The first family of approaches is based on computing a clustering-based similarity between nodes of the graph. One widely applied method is based on constructing a **consensus graph** out of the set of partitions to be combined [5,23]. The consensus graph G_{cons} is defined over the same set of nodes of the initial graph G. Two nodes $v_i, v_j \in V$ are linked in G_{cons} if there is at least one partition $P^y_{Q_x}$ where both nodes 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. Notice that the obtained graph is not necessarily a connected one. Different approaches can be applied in order to compute the aggregated clustering out from the consensus graph:

- In [23], authors transform the graph into a complete one by adding missing links with a null weight, then nodes are finally partitioned into clusters using agglomerative hierarchical clustering with some linkage rule, or by using a classical graph partitioning method such as the Kernighan-Lin algorithm [16].
- In [3] a similar approach is applied but with enforcing that nodes in the same result clusters should be connected in the initial graph by a sufficiently short path.
- In [20] authors propose a simple but effective method that consists on pruning links in the obtained consensus graph whose weights (frequency) is under a given threshold $\alpha \in [0, 1]$. The set of obtained connected components is taken to be the aggregated partition. The main problem of this approach is the problem of defining the value of the threshold α to use.

2.2 Ensemble Clustering-Based Community Detection

Ensemble clustering approaches have been used for various goals in the field of community detection in complex networks. One first direct application is to allow merging different partitions of the same graph obtained by applying a fast but low quality community detection algorithm, such as the label propagation algorithm [20]. Another application, to reduce the size of large-scale graphs. Let $G =< V, E >$ be a large)scale graphe. The idea is compute a set of n different low quality partitions of a graphe : $\Pi = \{\pi_1, \ldots, \pi_n\}$. A strict consigns graph is defined over the set of nodes V such that, $v \in V$ are linked if and only if they are grouped together in a same cluster in all partitions $\pi_i \in \Pi$. The obtained graph is usually composed of a large number of small connected components. Nodes composing each connected component are reduced to form only one node reducing hence the scale of the whole graph. The reduction phase can allow applying high quality community detection algorithms to the reduced graphe [22]. In [12] ensemble selection approaches are proposed in order to relaxe the constraint on connecting nodes if the frequency of being clustered together in all n partitions is higher than a given threshold $0 < \delta < 1$.

In [11], ensemble clustering approaches have been applied in order to implement multi-objective local community identification. In [10] an ensemble clustering approach is applied in order to compute a graph partition out of a set of bi-partitions of the graph computed after identifying local-communities of a set of seed nodes carefully selected to represent different points of view on the target graph.

Few work has addressed the problem of ensemble selection before applying the ensemble clustering process. In next section we introduce the problem of ensemble selection and we show how this can enhance the output of the ensemble clustering process.

3 Ensemble Selection

3.1 Problem Definition

Different works have showed that the quality of the output of an ensemble clustering is tightly related to both the *quality* of each partition in the base partitions set and *diversity* of these partitions.

Let Π be a set of n base clusterings. An ensemble selection function $\mathcal{ES}$ aims at selecting a subset of $\tilde{\Pi} \subseteq \Pi$ such that all partitions $\pi_i \in \tilde{\Pi}$ are of high quality and diverse. The diversity of partitions can be measured applying clustering comparison metrics such as the Adjusted Rand Index (ARI) [9], or information-based metrics such as the NMI [15].

The ARI index is based on counting the number of pairs of elements that are clustered in the same clusters in both compared partitions. Let $P_i = \{P_i^1, \ldots, P_i^l\}$, $P_j = \{P_j^1, \ldots, P_j^k\}$ be two partitions of a set of nodes V. The set of all (unordered) pairs of nodes of V can be partitioned into the following four disjoint sets:

- $S_{11} = \{$pairs that are in the same cluster under P_i and $P_j\}$
- $S_{00} = \{$pairs that are in different clusters under P_i and $P_j\}$
- $S_{10} = \{$pairs that are in the same cluster under P_i but in different ones under $P_j\}$
- $S_{01} = \{$pairs that are in different clusters under P_i but in the same under $P_j\}$

Let $n_{ab} = |S_{ab}|, a, b \in \{0, 1\}$, be the respective sizes of the above defined sets. The rand index, initially defined in [19] is simply given by :

$$\mathcal{R}(P_i, P_j) = \frac{2 \times (n_{11} + n_{00})}{n \times (n-1)}$$

In [9], authors show that the expected value of the Rand Index of two random partitions does not take a constant value (e.g. zero). They proposed an adjusted version which assumes a generalized hypergeometric distribution as null hypothesis: the two clusterings are drawn randomly with a fixed number of clusters and a fixed number of eleme nts in each cluster (the number of clusters in the two clusterings need not be the same). Then the adjusted Rand Index is the normalized difference of the Rand Index and its expected value under the null hypothesis. It is defined as follows:

$$ARI(P_i, P_j) = \frac{\sum\limits_{x=1}^{l} \sum\limits_{y=1}^{k} \binom{|P_i^x \cap P_j^y|}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \tag{2}$$

where:

$$t_1 = \sum_{x=1}^{l} \binom{|P_i^x|}{2}, t_2 = \sum_{y=1}^{k} \binom{|P_j^y|}{2}, t_3 = \frac{2t_1t_2}{n(n-1)}$$

This index has expected value zero for independent clusterings and maximum value 1 for identical clusterings.

Another family of partitions comparisons functions is the one based on the notion of mutual information. A partition P is assimilated to a random variable. We seek to quantify how much we reduce the uncertainty of the clustering of randomly picked element from V in a partition P_j if we know P_i. The Shanon's entropy of a partition P_i is given by:

$$H(P_i) = -\sum_{x=1}^{l} \frac{|P_i^x|}{n} log_2(\frac{|P_i^x|}{n})$$

Notice that $\frac{|P_i^x|}{n}$ is the probability that a randomly picked element from V be clustered in P_i^x. The mutual information between two random variables X,Y is given by the general formula:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) \quad (3)$$

This can then be applied to measure the mutual information between two partitions P_i, P_j. The mutual information defines a metric on the space of all clusterings and is bounded by the entropies of involved partitions. In [23], authors propose a normalized version given by:

$$NMI(X, Y) = \frac{MI(X, Y)}{\sqrt{H(X)H(Y)}} \quad (4)$$

The evaluation of the quality of a clustering is much harder, than the diversity, in unsupervised settings. In [1] authors propose to evaluate the quality of a partition $\pi_i \in \Pi$ by computing its distance (using ARI or NMI) from the consensus partition computed over the whole set Π. In [5], the quality of a partition $\pi_i \in \Pi$ us computed as follows: $Q(\pi) = \sum_{\pi \in \Pi} NMI(\pi, \pi_i)$.

In graph settings, external partition quality functions can be used to measure the equity of a partition. The well known *modularity* function is one option [8].

3.2 Proposed Approach

We propose here an original graph-based approach to cope with the problem of cluster ensemble selection. Algorithm 1 sketchs the general outlines of the proposed approach.

The algorithm is structured into four main steps. Having as an input a set of r base clusterings, we first compute an $r \times r$ pair-wise clustering similarity matrix M. An entry $M[i, j] = sim(r_i, r_j)$ gives the similarity between two base clusterings r_i and r_j. Different similarity functions can be used such as the

Algorithm 1. Graph-based cluster ensemble selection algorithm

```
Require: $G < V, E >$ a connected graph
Require: $\Pi = \{\pi_1, \ldots, \pi_r\}$ a base clusterings
Require: $\mathcal{Q}$ A partition quality function
$\Pi^* \leftarrow \emptyset$
$M \leftarrow$ compute_pairwise_similarity_Matrix($\Pi$)
GV ← construct_graph(M)
$\mathcal{C} = \{c_1, \ldots, c_k\} \leftarrow$ community_detection(GV)
for all $c \in \mathcal{C}$ do
    $\hat{\pi} \leftarrow \arg\max_{\pi \in c} Q(\pi)$
    $\Pi^* \leftarrow \Pi^* \cup \{\hat{\pi}\}$
end for
return $\Pi^*$
```

normalized mutual information (NMI), Adaptive Rand index (ARI) index or information variation (IV) [15,17]. The obtained matrix is then used to define a similarity graph GV over the set of base clusterings. Different kinds of similarity graphs can be defined. These include:

- **ϵ-Neighborhood Graph:** Here we connect all points whose pairwise distances are smaller than ϵ. As the distances between all connected points are roughly of the same scale ("at most"), weighting the edges would not incorporate more information about the data to the graph. Hence, the ϵ-neighborhood graph is usually considered as an unweighted graph.
- **k-Nearest Neighbor Graph:** Here the goal is to connect vertex v_i with vertex v_j if v_j is among the k-nearest neighbors of v_i. However, this definition leads to a directed graph, as the neighborhood relationship is not symmetric. There are two ways of making this graph undirected. The first way is to simply ignore the directions of the edges, that is we connect v_i and v_j with an undirected edge if v_i is among the k-nearest neighbors of v_j or if v_j is among the k-nearest neighbors of vi. The resulting graph is what is usually called the k-nearest neighbor graph. The second choice is to connect vertices v_i and v_j if both v_i is among the k-nearest neighbors of v_j and v_j is among the k-nearest neighbors of v_i. The resulting graph is called the mutual k-nearest neighbor graph. In both cases, after connecting the appropriate vertices we weight the edges by the similarity of their endpoints.
- **Relative Neighborhood Graph:** Relative neighborhood graph (RNG) has been initially proposed in [25]. The choice of RNG graph is motivated by the topological characteristics of these graphs that are connexe and sparse. To build an RNG graph, we first compute a similarity matrix between couple of items in the dataset. This results in a symmetric square matrix of size $n \times n$ where n is the number of items in the dataset. A RNG graph is defined by the following simple construction rule: two points x_i and x_j are connected by an edge if they satisfy the following property:

$$d(x_i, x_j) \leq \max_{l}\{d(x_i, x_l), d(x_j, x_l)\}, \forall l \neq i, j \tag{5}$$

where $d(x_i, x_j)$ is the distance function. A community detection algorithm is applied on the obtained graph in order to cluster the given examples. Clustering evaluation criteria can then be used to compare different algorithms.

In this work, we have selected to build a relative neighborhood graph since it is the only approach that guarantee having a connected and sparse graph.

4 Experiments

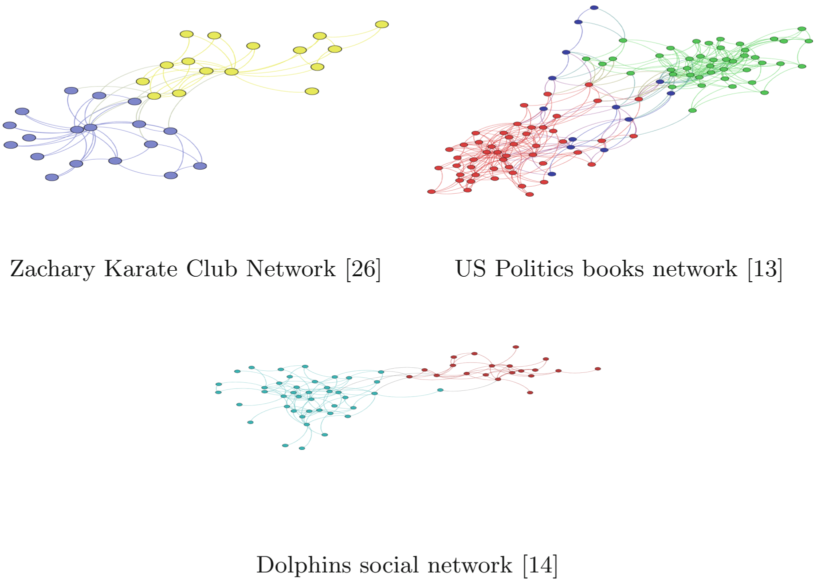
In this section we evaluate the utility of the proposed ensemble selection approach for enhancing community detection in real world complex networks. The evaluation process is the following: given a network for which we know a ground truth partition into communities we apply first the label propagation approach 100 times. We then compute a consensus partition applying a CSPA ensemble clustering approach on the whole set of obtained partitions and on the set of partitions selected by applying our approach. The quality of obtained communities is evaluated using the ARI and NMI metrics with respect to the ground truth partition.

A set of three widely used benchmark networks for which a ground-truth decomposition into communities are known are used. These are the following:

- **Zachary's Karate Club:** This network is a social network of friendships between 34 members of a karate club at a US university in 1970 [26]. Following a dispute the network was divided into 2 groups between the club's administrator and the club's instructor. The dispute ended in the instructor creating his own club and taking about half of the initial club with him. The network can hence be divided into two main communities.
- **Dolphins Social Network:** This network is an undirected social network resulting from observations of a community of 62 dolphins over a period of 7 years [14]. Nodes represent dolphins and edges represent frequent associations between dolphin pairs occurring more often than expected by chance. Analysis of the data revealed two main groups.
- **American Political Books:** This is a political books co-purchasing network. Nodes represent books about US politics sold by the online bookseller *Amazon.com*. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon. Books are classified into three disjoint classes: liberal, neutral or conservative. The classification was made separately by Mark Newman based on a reading of the descriptions and reviews of the books posted on Amazon.

Next figure shows the structure of the selected networks with real communities indicated by the color code. In Table 1 we summarize basic characteristics of selected benchmark real networks (Fig. 1).

For all three datasets, the ensemble selection process enhance the quality of the obtained final partition (Table 2).


Zachary Karate Club Network [26]

US Politics books network [13]

Dolphins social network [14]

Fig. 1. Real community structure of the selected benchmark networks

Table 1. Characteristics of some well-known benchmark networks

Network	# nodes	# edges	# com	reference
Zachary club	34	78	2	[26]
Political books	100	441	3	[13]
Dolphins	62	159	2	[14]

Table 2. Evaluation if he proposed graph-based ensemble selection

Dataset	Approach	NMI	ARI	Q	# Communities
Zachary	Ensemble clustering without selection	0.57	0.46	0.40	5
	Ensemble clustering with selection	**0.77**	0.69	0.34	2
US Politics	Ensemble clustering without selection	0.55	0.68	0.51	5
	Ensemble clustering with selection	**0.68**	0.67	0.42	6
Dolphins	Ensemble clustering without selection	0.55	0.39	0.51	5
	Ensemble clustering with selection	**0.58**	0.59	0.53	3

5 Conclusion

Ensemble clustering approaches are proposed as mean to cope with the robustness issue. of high speed community detection algorithms. In this work, we have proposed a new approach for enhancing the output of ensemble clustering by applying an ensemble selection process. An original graph-based ensemble selection approach is studied. Results show that the overall quality of detected communities 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

1. Azimi, J., Fern, X.: Adaptive cluster ensemble selection. In: Boutilier, C. (ed.) IJCAI, pp. 992–997 (2009)
2. Blondel, V.D., Guillaume, J.I., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**, P10008 (2008)
3. Dahlin, J., Svenson, P.: Ensemble approaches for improving community detection methods. CoRR abs/1309.0242 (2013)
4. Fern, X.Z., Lin, W.: Cluster ensemble selection. Stat. Anal. Data Min. **1**(3), 128–141 (2008)
5. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: Brodley, C.E. (ed.) ICML. ACM International Conference Proceeding Series, vol. 69. ACM (2004)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
7. Fred, A.L.N., Jain, A.K.: Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 835–850 (2005)
8. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS **99**(12), 7821–7826 (2002)
9. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 192–218 (1985)
10. Kanawati, R.: YASCA: an ensemble-based approach for community detection in complex networks. In: Cai, Z., Zelikovsky, A., Bourgeois, A. (eds.) COCOON 2014. LNCS, vol. 8591, pp. 657–666. Springer, Heidelberg (2014)
11. Kanawati, R.: Empirical evaluation of applying ensemble methods to ego-centered community identification in complex networks. Neurocomputing **150, B**, 417–427 (2015)
12. Kanawati, R.: Ensemble selection for enhancing graph coarsening quality. In: Proceedings of 5th International Workshop on Social Network Analysis. Capri, April 2015
13. Krebs, V.: Political books network. http://www.orgnet.com
14. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav. Ecol. Sociobiol. **54**, 396–405 (2003)
15. Meila, M.: Comparing clusterings by the variation of information. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 173–187. Springer, Heidelberg (2003)
16. Newman, M.: Networks: An Introduction. Oxford University Press, Oxford (2010)
17. Nguyen, X.V., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
18. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 1–12 (2007)

19. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
20. Seifi, M.: Cœurs stables de communautés dans les graphes de terrain. Ph.D. thesis, Université Pierre et marie Curie (paris 6) (2012)
21. Seifi, M., Guillaume, J.L.: Community cores in evolving networks. In: Mille, A., Gandon, F.L., Misselis, J., Rabinovich, M., Staab, S. (eds.) WWW (Companion Volume), pp. 1173–1180. ACM, New York (2012)
22. Staudt, C., Meyerhenke, H.: Engineering high-performance community detection heuristics for massive graphs. In: ICPP, pp. 180–189. IEEE (2013)
23. Strehl, A., Ghosh, J.: Cluster ensembles: a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
24. Tang, L., Liu, H.: Community Detection and Mining in Social Media. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, San Rafael (2010)
25. Toussaint, G., Bhattacharya, B.K.: Optimal algorithms for computing the minimum distance between two finite planar sets. Pattern Recogn. Lett. **2**(2), 79–82 (1981)
26. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)