Evidential community detection using structural and attribute information

Kuang Zhou, Arnaud Martin, Quan Pan

Evidential community detection using structural and attribute information

Kuang Zhou [1,2], **Arnaud Martin** [2], **Quan Pan** [1]

1. *School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, PR China*
   kzhoumath@163.com
2. *DRUID, IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France*
   Arnaud.Martin@univ-rennes1.fr

RÉSUMÉ. *L'objectif de la détection de communautés est de créer une partition des sommets, de telle sorte que les communautés soient composées de sommets fortement connectés. Les approches existantes de détection de communautés se concentrent principalement sur la structure topologique du réseau, mais elles ignorent largement les informations disponibles à propos des attributs des nœuds. Dans cet article, une nouvelle appproche de détection de communautés qui utilise à la fois les informations structurelles et d'attributs pour extraire une structure de graphe imprécise, est proposée dans le cadre de la théorie des fonctions de croyance. L'objectif de notre méthode consiste à partitionner les sommets dans différents groupes afin que chaque cluster contienne un sous-graphe connecté densément avec les valeurs de l'attribut homogène. Les résultats expérimentaux montrent l'efficacité de la méthode proposée et montrent qu'elle pourrait améliorer les performances de la détection de communautés où les informations sur les propriétés du graphe sont disponibles en complément avec la structure topologique.*

ABSTRACT. *The goal of community detection is to partition nodes into different small subgroups in such a way that vertices in the same community have strong connections. Existing community detection approaches mainly focus on the topological structure of the network, but ignore the information about node attributes. In this paper, a new Evidential Community detection approach which could utilize both Structural and Attribute information, named ECSA, is proposed using belief functions to extract imprecise graph structure. The goal of our method is to partition vertices into different groups so that each cluster contains a densely connected subgraph with homogeneous attribute values. Experimental results illustrate the effectiveness of the proposed method and show that it could indeed improve the performance of community detection when the information about vertex properties is available together with topological structure.*

MOTS-CLÉS : *Détection de communautés; Attributs des nœuds; Communautés imprécises; Fonctions de croyance.*

KEYWORDS: *Community detection; Node attributes; Imprecise communities; Belief functions.*

DOI:10.3166/RIA. .1-**??**

1. Introduction

Community detection is a useful unsupervised learning technique for detecting the cohesive groups in networks, and it has been developed rapidly in recent years and widely used in many applications. Traditional community detection approaches are mostly based on the topological structure of the graph. Sometimes, the vertex properties could also provide valuable information to guide the community detection process. For example, in a social network such as Facebook, users may have the following attributes: ID, a student/faculty status flag, gender, major, high school and college. Besides, traditional methods mostly focus on the non-overlapping communities. The work of detection overlapping communities incorporating both structural and attribute information has not been thoroughly studied yet. This is the motivation of our work.

In this paper, a new Evidential Community detection approach using both graph Structure and node Attributes (ECSA) is proposed in the framework of belief functions. The approach is based on an enhanced version of Evidential C-Means (ECM). An item is added into the objective function of ECM to reflect the consistence of members in the same group in terms of attribute information. The best partitions of the networks are obtained by optimizing the objective function. The experimental results demonstrate that the proposed approach could improve the performance of community detection when multiple information about the graph is available.

2. Related work

Detecting communities is still an open problem in social network analysis. Recently, significant progress has been achieved in this research field and several popular algorithms for community detection have been put forward. Among them we mention the modularity-based methods (Newman, 2006), label propagation algorithm (Raghavan *et al*., 2007), spectral optimization method (White, Smyth, 2005), and see (Fortunato, 2010) for review of the topic. However, these algorithms ignore the attributes of the nodes. Several new clustering methods that use both structure and attributes of graphs are introduced in recent years, such as SA-Cluster (Cheng *et al*., 2011), where a unified distance measure to combine structural and attribute similarities is defined, and then a clustering strategy similar to k-medoids is adopted to partition the nodes. Some other methods can be seen in the work of (Yang *et al*., 2013), (Ge *et al*., 2008), (Xu *et al*., 2012) and so on.

Recently, (Masson, Denoeux, 2008) proposed the application of evidential c-means (ECM) to get credal partitions for object data. The credal partition is a general extension of the crisp (hard) and fuzzy ones and it allows the object to belong to not only single clusters, but also any subsets of the set of clusters $\Omega = \{\omega_1, \cdots, \omega_c\}$ by allocating a mass of belief for each object in X over the power set 2^{Ω}. The additional flexibility brought by the power set provides more refined partitioning results than those by the other techniques allowing us to gain a deeper insight into the data.

3. Proposed method

Here we present the proposed ECSA algorithm in detail. ECSA is based on an enhanced version of ECM with Relational information (ECMwR). Therefore we will describe ECMwR first and then give the detailed ECSA algorithm.

3.1. ECMwR clustering

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ be a collection of vectors in $\mathbb{R}^p$ describing n objects to be classified into c clusters in the set $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$. Assume that some proximity information about the data set given in the form of relation matrix $\boldsymbol{W}_{n\times n}$ is available, and $w_{ij} \in \boldsymbol{W}$ represents the relationship between object $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Here we let $w_{ii} = 0, \forall i = 1, 2, \cdots, n$. The objective function of ECMwR is given as below:

$$\begin{aligned} J_{\text{ECMwR}} &= J_{\text{ECM}} + J_{\text{Re}} \\ &= \sum_{i=1}^{n} \sum_{A_h \subseteq \Omega, A_h \neq \emptyset} |A_h|^{\alpha} \, m_i^2(A_h) d_{ih}^2 + \sum_{i=1}^{n} \delta^2 m_i^2(\emptyset) + \tau \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \mathcal{K}_{ij} \end{aligned} \tag{1}$$

with

$$\mathcal{K}_{ij} = \begin{cases} \sum\limits_{A_e \cap A_r = \emptyset} m_i\left(A_e\right) m_j\left(A_r\right) & i \neq j, \\ 0 & i = j. \end{cases} \tag{2}$$

The term $\mathcal{K}_{ij}$, which in fact is the mass assigned to the empty set by the conjunctive combination, reflects the disagreement between masses for $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Parameter α is a tuning parameter allowing to control the degree of penalization for subsets with high cardinality, and d_{ih} denotes the distance (generally Euclidean distance) between x_i and the barycenter (*i.e.* prototype) associated with A_h, and τ balances the contribution of two components, *i.e.,* J_{ECM} and J_{Re}. Parameter δ is used to detect outliers. The objective function of J_{ECMwR} should be subject to constraints in Eq. (3).

$$\sum_{A_j \subseteq \Omega, A_j \neq \emptyset} m_i(A_j) + m_i(\emptyset) = 1, m_i\left(A_j\right) \geq 0, \;\; m_i\left(\emptyset\right) \geq 0. \tag{3}$$

To solve the constrained minimization problem, the method of Lagrange multipliers provides a classical way. An alternate optimization scheme similar to that in FCM and ECM algorithms can be designed for ECMwR with this method. First, we consider that the prototype set of clusters, V, is fixed. The update equations of $m_{ij} \triangleq m_i(A_j)$ for ECMwR could be derived as follows.

$$m_{ik} = m_{ik}^{\text{ECM}} + \tau m_{ik}^{\text{Re}} \tag{4}$$

with

$$m_{ik}^{\text{ECM}} = \frac{|A_k|^{-\alpha} d_{ik}^{-2}}{\sum\limits_{A_h \subseteq \Omega, A_h \neq \emptyset} |A_h|^{-\alpha} d_{ih}^{-2} + \delta^{-2}}, \;\; \forall i = 1, 2 \cdots, n, \forall k / A_k \subseteq \Omega, A_k \neq \emptyset, \tag{5}$$

$$m_{ik}^{\text{Re}} = \frac{\frac{\sum\limits_{A_h \subseteq \Omega, A_h \neq \emptyset} \sum_{j=1}^{n} w_{ij} \sum\limits_{A_h \cap A_l = \emptyset} m_{jl} |A_h|^{-\alpha} d_{ih}^{-2} + \sum\limits_{j=1}^{n} w_{ij} \delta^{-2}}{\sum\limits_{A_h \subseteq \Omega, A_h \neq \emptyset} |A_h|^{-\alpha} d_{ih}^{-2} + \delta^{-2}} - \sum_{j=1}^{n} w_{ij} \sum\limits_{A_k \cap A_l = \emptyset} m_{jl}}{|A_k|^{\alpha} d_{ik}^2}, \tag{6}$$

and

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij}, \ \ \forall i = 1, 2, \cdots, n. \tag{7}$$

It is remarkable that Eq. (4) is a group of equations of m_{ik}, and the constraints are not explicitly satisfied. To obtain the solution of Eq. (4), the simple successive-substitution method, in which one can repeatedly use old values of m_{ik} in Eq. (6) to get m_{ik}^{Re} and then solve for new values of m_{ik} from Eq. (4) until convergence, could be utilized. In practice, one can improve the order of convergence of this approach by the application of Seidel iteration scheme, where all the new available mass values are used for solving m_{ik}.

As we can see, the update formula of m_{ik} derived in Eq. (4) does not guarantee non-negativity. The Karush–Kuhn–Tucker (KKT) conditions could be used to force the memberships to be positive. Since the application of KKT conditions would yield more complicated update equations, in this paper we use a simple clipping strategy: at each iteration set those negative mass values obtained by Eq. (4) to 0, and renormalize the values to sum to 1.

From Eq. (1) it is easy to know the penalty term J_{Re} in the objective function of ECMwR does not depend on the cluster centroids. Thus, given the partition matrix $\boldsymbol{M}$, the updating of the prototype set $\boldsymbol{V}$ in ECMwR could be evoked by the same scheme as that in the application of ECM (Masson, Denoeux, 2008).

3.2. Community detection approach

An attributed graph is denoted as $G = (V, E, A)$, where V is the set of nodes, E is the set of edges, and $A = \{a^1, a^2, \cdots, a^d\}$ is the set of d attributes associated with nodes in V. Each vertex v_i is associated with an attribute vector $\left(a_i^1, \cdots, a_i^d\right)$. ECSA algorithm is described in the following.

(1) Map the topological structure to feature vectors by some spectral methods;

(2) Construct the relational matrix based on the attribute values; In this work we only consider discrete attributes. The similarity between two nodes could be determined by examining each of d attributes and counting the number of attribute values they share in common;

(3) Evoke ECMwR clustering algorithm and obtain the detected communities.

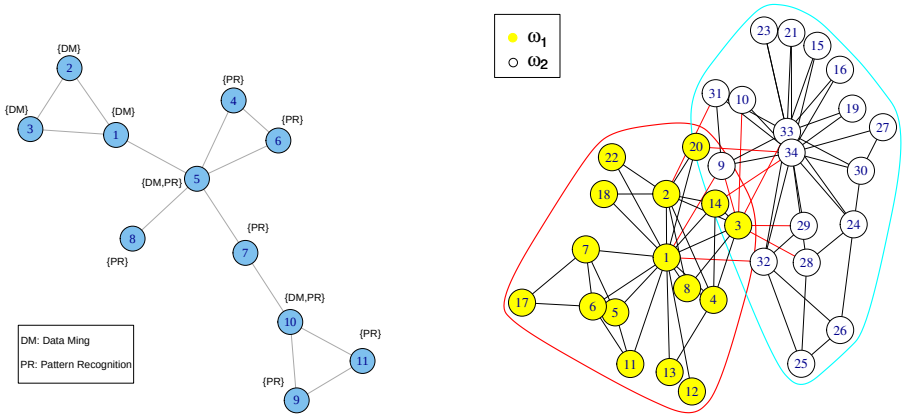
Remark: As in ECM, in ECMwR the number of parameters to be optimized is exponential and depends on the number of clusters. For the number of classes larger than 10, calculations are not tractable. But we can consider only a subclass with a limited number of focal sets. For instance, we could constrain the focal sets to be composed of at most two specific classes.

4. Experiments

To show the principle of the proposed method, a small illustrative example of a co-authorship network displayed in Figure 1.a is first considered. In the graph each vertex represents an author while an edge represents the co-author relationship between two

authors. In addition, there are primary topics associated with each author. The research topic is regarded as an attribute describing the vertex property. The graph structure is mapped into Euclidian space using signal prorogation method (Hu *et al.*, 2008) and the vertex attributes are used to construct the relation matrix w_{ij}. For the author i and j, if they share r same topics, $w_{ij} = r$; Otherwise $w_{ij} = 0$.

Two other community detection schemes are compared. One is FCM based clustering after the spectral mapping, the other is ECM based clustering. Note that in these two algorithms, only the graph structure is used. The authors are assigned to the groups with maximal mass value by ECM and ECMwR. The results show that by FCM and ECM the authors in the same detected group by FCM and ECM may have different topics. On the contrary, authors in the same found community by ECMwR are not only closely connected, but also sharing homogeneous research topics. Credal partitions enable us to detect the imprecise overlapped nodes. The overlapped node found by ECM is node 7, while node 5 is regarded as overlapping by ECMwR. Author 5 and Author 10 regard both DM (Data Mining) and PR (Pattern recognition) as their own topics, but only Author 5 is clustered into the imprecise cluster by ECMwR. This is due to the fact that Author 5 has collaborated with authors in both two communities while Author 10 only has co-authorship with those whose topic is PR. The detection result by ECMwR is more reasonable as it takes advantage of the structural and attribute similarities at the same time.



a. Co-Authorship Network b. Karate Club Network

Figure 1. Original networks.

In the next experiment we will use a widely used benchmark in detecting community structures, “Karate Club", studied by Wayne Zachary. The network consists of 34 nodes and 78 edges representing the friendship among the members of the club (see Figure 1.b). In the original network there is no attribute for nodes. We generate random attributes (a_{i1}, a_{i2}) for node i based on the ground-truth. For all the nodes in community ω_1, $a_{i1} = 1, a_{i2} = 0$, while for those in ω_2, $a_{i1} = 0, a_{i2} = 1$. Then we modify the attribute of node 3 to $(1, 1)$. From the results we see, without the attribute information, nodes 3, 9, 14 are all partitioned into the imprecise class $\{\omega_1, \omega_2\}$ by

ECM. But after taking the attribute information into account, only node 3 is regarded as a member in $\{\omega_1, \omega_2\}$. This is due to the fact that node 3 lies in the overlap in terms of not only topological structure, but also attribute values.

From the two experiments, we can get that: 1. The found communities by ECSA contains members not only being connected closely but also sharing similar attributes. 2. ECSA could detect overlapping communities in the concept of credal partitions. Also for the members in the overlap, they are not only frequently connected to the vertices in all the related groups but also labelled with more than one attribute.

5. Conclusion

In this study, an evidential community detection method incorporating both structural and attribute information is presented in the framework of belief functions. The proposed algorithm is based on an enhanced version of ECM clustering using available relational information. Experimental results show that our method will provide detected communities with nodes not only being densely connected but also have homogeneous attribute values.

Bibliographie

Cheng H., Zhou Y., Yu J. X. (2011). Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, n° 2, p. 12.

Fortunato S. (2010). Community detection in graphs. *Physics Reports*, vol. 486, n° 3, p. 75–174.

Ge R., Ester M., Gao B. J., Hu Z., Bhattacharya B., Ben-Moshe B. (2008). Joint cluster analysis of attribute data and relationship data: The connected k–center problem, algorithms and applications. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, n° 2, p. 7.

Hu Y., Li M., Zhang P., Fan Y., Di Z. (2008). Community detection by signaling on complex networks. *Physical Review E*, vol. 78, n° 1, p. 016115.

Masson M.-H., Denoeux T. (2008). Ecm: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, vol. 41, n° 4, p. 1384–1397.

Newman M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, vol. 103, n° 23, p. 8577–8582.

Raghavan U. N., Albert R., Kumara S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, vol. 76, n° 3, p. 036106.

White S., Smyth P. (2005). A spectral clustering approach to finding communities in graph. In *Sdm*, vol. 5, p. 76–84.

Xu Z., Ke Y., Wang Y., Cheng H., Cheng J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, p. 505–516.

Yang J., McAuley J., Leskovec J. (2013). Community detection in networks with node attributes. In *Data mining (ICDM), 2013 ieee 13th international conference on*, p. 1151–1156.