Complex Network Community Extraction Based on Gaussian Mixture Model Algorithm

Complex Network Community Extraction Based on Gaussian Mixture Model Algorithm

Dai Ting-ting[1], **Dong Yan -shou**', **Shan Chang-ji** [2]

([1] School of mathematics and statistics, Zhaotong University, Yunnan, 657000, China,

[2] School of physics and electronic information engineering, Zhaotong University, Yunnan, 657000, China)

Abstract: Based on the problem of community partitioning in complex networks,this paper proposes a Gaussian mixture model community extraction algorithm based on principal component analysis.The idea of the algorithm is as follows:Firstly,the principal component analysis is used to reduce the dimension of the adjacency matrix of the network;secondly,it is assumed that the communities in a network are generated by different Gaussian models,that is,the generation mechanism of different models is different;The parameters of the model are solved by the expectation maximization algorithm. Simulation experiments show that if the contribution rate of the principal component reaches more than 90%, the network division is very consistent with the actual network,and the time used is also short. Compared with other methods,it has obvious advantages.

1. Introduction

Through the continuous research of experts and scholars in various fields,the algorithm for extracting community structure in complex networks has achieved certain achievements at home and abroad.The most classic algorithm is based on graphics segmentation in computer science.Kerjnighan-Lin algorithm [1] and spectral bisection algorithm [2]; clustering GN algorithm [3]; later many researchers proposed a series of new algorithms based on GN algorithm,the purpose is to ensure accuracy Under the premise,and it can reduce the time complexity of the algorithm running. Although some classical community extraction algorithms have been obtained through the efforts of scholars on the basis of the original algorithms,the application scope of these algorithms is still small, and it cannot solve most community extraction problems.In addition,these algorithms cannot balance the accuracy and time algorithm complexity, and cannot meet the requirements of practical applications. To this end, in recent years,researchers have designed some community extraction algorithms based on statistical theory reasoning.Such algorithms can not only identify the generalized community structure in the network,but also identify the characteristics of structural equivalence and regular equivalence in the network.Such methods not only have a solid theoretical basis,but also have high partitioning accuracy while the computational time complexity is low.This paper proposes a community structure extraction algorithm based on this theory---the Gaussian mixture model community extraction algorithm based on principal component analysis.

2. Method and theory

2.1 Gaussian mixture model

This paper introduces a popular clustering algorithm----the node of Gaussian mixture model [4]. Under this model,we think that data $X=\{x^{(1)},x^{(2)}...x^{(N)}\}$ is composed of multiple Gaussian models,and any Gaussian distribution $N(x;\mu_j,\Sigma_j)$ Is a cluster center. The difference between this clustering method and the $k-means$ algorithm is the introduction of probability in the Gaussian mixture model. Therefore,in the process of classification,a probability distribution of the value of Y is obtained by the data X,that is,the output result obtained through continuous learning and training is not a specific value,but a probability value of a series of values,which can be based on these probabilities A comprehensive analysis of the observed objects is performed using different models to make judgments and decisions.

There are N observation sample data $X=\{x^{(1)},x^{(2)}...x^{(N)}\}$, assuming that these observation data are generated by multiple Gaussian models.If each Gaussian model obeys the $N(\mu_{j,}\Sigma_j)$ distribution, then there are $X\sim\sum_{j=1}^{K}\phi_j N(\mu_j,\Sigma_j)$,and $\sum_{j=1}^{K}\phi_j=1$,where k indicates that the model is a mixture of k Gaussian models, φ_j is the weight coefficient of each model,and μ_j and Σ_j represent the mean of the Gaussian distribution,respectively.and the variance,the comprehensive Gaussian mixture model can be obtained as:

$$p(x,\theta)=\sum_{j=1}^{k}\varphi_j f(x;\mu_j,\Sigma_j) \tag{1}$$

Where $f(x;\mu_j,\Sigma_j)=\frac{1}{\sqrt{(2\pi)^d\left|\Sigma_j\right|}}\exp[-\frac{1}{2}(x-\mu_j)^T\Sigma_j{}^{-1}(x-\mu_j)]$,$d$ is the dimension of the observed sample and θ is the set of parameter vectors.

2.2 Expectation maximization algorithm

Knowing the sample points,but not knowing the sample classification,that is,containing the hidden variables,we need to calculate the parameters ϕ,μ and Σ in the model (1).The usually effective solution method is Expectation Maximization (EM). Algorithm [5].

According to the literature [6],the EM algorithm is actually an iterative algorithm consisting of two steps (E-Step and M-Step).In the problem of this paper, E-Step is the constant "guess" the value of $z^{(i)}$,in M- In the Step,the parameter values of the model are continuously updated according to the value of the guessed $z^{(i)}$,and the above steps are repeated until convergence.The algorithm is as follows:

E-Step: For each i,j setting

$$w_j{}^{(i)}=p(z^{(i)}=j\left|x^{(i)};\phi,\mu,\Sigma\right.)$$

$$=\frac{p(x^{(i)}\left|z^{(i)}=j\right.)p(z^{(i)}=j)}{\sum_{l=1}^{k}p(x^{(i)}\left|z^{(i)}=l\right.)p(z_l{}^{(i)}=l)} \qquad (2)$$

$$=\frac{p(x^{(i)}\left|z^{(i)}=j;\mu,\Sigma\right.)p(z^{(i)}=j;\varphi)}{\sum_{l=1}^{k}p(x^{(i)}\left|z^{(i)}=l;\mu,\Sigma\right.)p(z^{(i)}=l;\varphi)}$$

In E-Step,the value of the posterior probability $w_j{}^{(i)}$ of the parameter $z^{(i)}$ is calculated from the values of $x^{(i)}$ and the current parameters ϕ,μ,Σ to represent our "soft estimate" of $z^{(i)}$.

M-Step: Update the following parameters based on (2.2-1):

$$\varphi_j = \frac{1}{N}\sum_{i=1}^{N}\omega_j^{(i)} \tag{3}$$

$$\mu_j = \frac{\sum_{i=1}^{N}\omega_j^{(i)}x^{(i)}}{\sum_{i=1}^{N}\omega_j^{(i)}} \tag{4}$$

$$\Sigma_j = \frac{\sum_{i=1}\omega_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{N}\omega_j^{(i)}} \tag{5}$$

Where $\omega_j^{(i)}$ is the probability that the sample point i belongs to class j,not the indicator function that sample point i belongs to class j.And in this mixed model, we use different covariance matrices Σ_j.

2.3 for Expectation maximization algorithm in Gaussian mixture model different models.

According to the derivation formula of the EM algorithm [7] in the Gaussian mixture model,we obtain the following steps to solve the parameters of the Gaussian mixture model by EM algorithm:

Step 1: Initialization parameters, randomly define the mean μ_j, ϕ_j take $\frac{2}{N}$, and the covariance matrix e takes the unit matrix.

Step 2: In E-Step,according to the calculation method of $\omega_j^{(i)}$ in (3),the calculation node i is the probability $\omega_j^{(i)}$ generated by the j th Gaussian model,that is,the probability that the node i belongs to the jth community.

Step 3: That is,M-Step,on the basis of the second step,update the values of φ_j, μ_j and Σ_j according to (3), (4), (5), respectively.

Step 4: The loop repeats steps 2 and 3,and the value of the parameter θ is continuously updated. When the difference between the maximum likelihood function values obtained twice after the current time is less than the critical value,the algorithm stops.

2.4 Principal component analysis

We make the Gaussian mixture model that can be used to deal with large complex networks, according to the degree of relevance between nodes of complex networks,this paper introduces Principal Componet Analysis (PCA) [8],which is applied in the extraction of complex networks. Principal component analysis is the main innovation of this paper,and it is also the main means to shorten the running time of community extraction algorithm. Since Principal Component Analysis has no parameter limitation in practical applications,it is widely used as a dimension reduction tool.When processing high-dimensional data,it can undergo a series of changes in the pre-processing stage to high-dimensional space.The data is mapped to a low dimensional space.We remove noise redundancy data and find the part that best represents the original data.

In dealing with big data,we need to know the relationship between the dimensions of the sample. The problem of reaction is the covariance matrix of the sample.This matrix is a pair of matrices,and the elements in it represent the covariance between the dimensions.The purpose of introducing principal component analysis is to make low-dimensional numbers so that the remaining data can represent the original data,that is,the correlation between different latitudes is required to be weakest after dimension reduction,that is,the non-diagonal elements of the covariance matrix are as zero as possible.Then,we can diagonalize the covariance matrix,which completes the step of denoising.This shows that the purpose of principal component analysis is to diagonalize the covariance matrix of the sample.

2.5 Gaussian mixture model based on principal component analysis

In the previous community extraction algorithm,the adjacency matrix of the network is directly used for calculation.When the number of nodes in the network is relatively large,the connection between the network nodes is complicated,which leads to high complexity of the algorithm and time consuming.Too long, but also limits the existing algorithms can not handle large complex networks, such as biological networks,the Internet,social networks,and so on.Since the degrees between nodes of a complex network are related,specifically,the variables in each row (or column) of the adjacency matrix are not isolated,then all variables can be replaced with as few variables as possible.Based on this, the principal component analysis can be used to reduce the dimension of the column vector of the adjacency matrix of the network before the algorithm is performed,so that the time complexity of the subsequent calculation can be reduced.

Let the adjacency matrix of the network be A,N is the total number of nodes in the network,and each row of the matrix A corresponds to one node.If there are edges between i and j,then y,otherwise r, if the network is an undirected network,then d is a symmetric matrix.The steps of the Gaussian mixture model extraction community algorithm based on principal component analysis proposed in this paper are as follows:

Step 1: For the normalization process of the adjacency matrix A,the covariance matrix is calculated according to the formula (3);

Step 2: the eigenvalue of the covariance matrix C and the contribution rate of each eigenvalue of the eigenvector are calculated ,we let $\lambda_1,\lambda_2,...,\lambda_N$ be the N eigenvalues of C; $\tau_1,\tau_2,...,\tau_N$ be the corresponding eigenvector,and arrange the eigenvalues according to the eigenvalues from large to small: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$,the eigenvalue contribution rate and the cumulative contribution rate are defined: $\lambda_i / \sum_{i=1}^{N} \lambda_i$ and $\sum_{i=1}^{p} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i$;

Step 3: Pre-*P(P<N)* principal components are extracted according to the principle of cumulative contribution rate $\sum_{i=1}^{p} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i$, that is,*P* mutually orthogonal eigenvector matrices $A' = (\tau_1, \tau_2, ...\tau_p)$ corresponding to *P* eigenvalues are retained;

Step 4: Using the eigenvector matrix A' to linearly transform $A'' = AA'$ the network adjacency matrix A to obtain the low-dimensional matrix A'' of the adjacency matrix,the dimensionality reduction operation of the adjacency matrix is completed,and the observation data of the Gaussian mixture model is A'';

Step 5: Initialize the parameters ϕ_j , μ_j and Σ_j values,randomly define the mean μ_j , ϕ_j take $\frac{2}{N}$,and the covariance matrix Σ_j takes the identity matrix;

Step 6: According to the $\omega_j^{(i)}$ calculation method in (2),the calculation node i is the probability $\omega_j^{(i)}$ generated by the jth Gaussian model,that is,the probability that the node i belongs to the j-th society;

Step 7: On the basis of step 6,the values of μ_j , ϕ_j and Σ_j are respectively calculated according to the formulas (3)(4)(5);

Step 8: Repeat the loop operation steps 6,7 and continuously update the value of the parameter θ. When the difference between the maximum likelihood function values obtained twice after the current time is less than the critical value,the algorithm stops.

Obtain a series of $\omega_j^{(i)}$,that is,the "softening point" result of the network, and select the class with the largest $\omega_j^{(i)}$ value as the class label of node i ,which realizes the division of the network.

3. Empirical analysis

3. 1 Zachary network[8] *results analysis*

Since a known network is composed of two communities,the set hybrid model (1) is a mixture of two Gaussian models. Using (2), (3), (4), and (5) pairs,a series of parameters ϕ μ, Σ , $\omega_j^{(i)}$ in the model are solved,and the obtained $\omega_j^{(i)}$ value is the "softening point" result of the network,that is,the probability that the node belongs to the community,the red square The box indicates the probability that the node belongs to the community 2,and the class with the highest attribution probability is selected as the community label of the node.The resulting network partitioning results are shown in Figure1.By comparing with the actual division of the network,we find that the results of the text division are exactly the same as the actual division results.According to the formula (5),the calculated value is1.This high accuracy is not achieved by many existing methods.Therefore,we speculate that the generation mechanism of the network is consistent with the Gaussian mixture model.

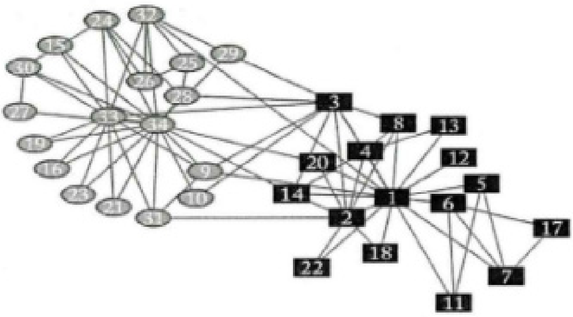

Figure 1 Schematic diagram of Zachary Karate Club network division results

In order to verify that principal component analysis is introduced into the complex network community extraction,it is effective to increase the number of principal components in the Zachary network, increase the contribution rate of the principal component,and then observe the accuracy of the network segmentation results to obtain the principal component contribution rate. Figure 2. NMI value.The diamond in the figure represents the cumulative contribution rate of the principal component,and the square represents the corresponding NMI value. It can be seen from the figure that when the cumulative contribution rate of the principal component increases,the NMI value increases in a stepwise manner.In particular,when the number of principal components reaches 14,and the cumulative contribution rate reaches 0.945 or more,the network can be completely and accurately divided.

Relationship between cumulative contribution rate of principal component and NMI value

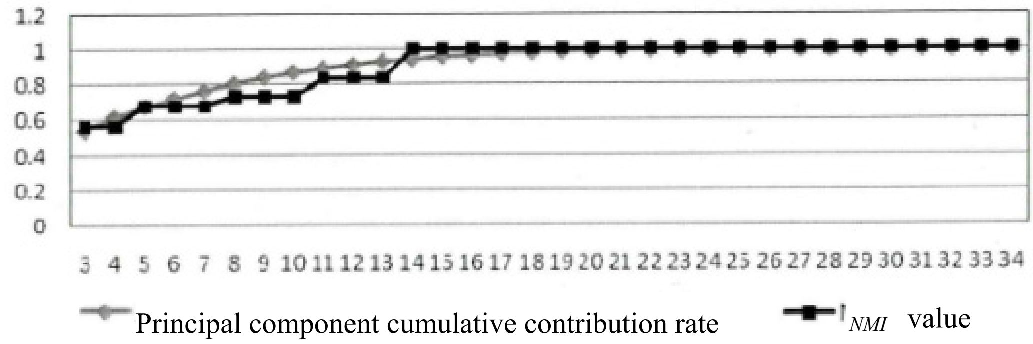

Figure 2 The relationship between the cumulative contribution rate of the principal components and the NMI,the abscissa indicates the number of principal components,and the ordinate indicates the NMI value and the cumulative contribution rate of the principal components

4. Conclusion

Based on the network dataset of Zachary Karate Club,this paper proposes a Gaussian mixture model community extraction algorithm. The experimental results show that compared with the previous algorithms,the algorithm not only greatly improves the accuracy,but also greatly reduces the running time of the algorithm..

References

[1]Kermighan B W,Lin S An efficient. Heuristic procedure for partitioning graphs[J].Bell system technical Journal 1970,49(2):291-307.

[2]Newman M.E.J.modularity and community Structure in networks[J].proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.

[3]Giran M, Newman M .E .J. community structure in social and biological networks[J].proceedings of the National Academy of Sciences2002,99(12):7821-7826

[4]Redner R A,Walker HF.Mixture densities,maximum likelihood and the EM,algorithm[J].SIAM review,1984,26(2):195-239.

[5]Dempster AP,Laird NM,Rubin D B. Maximun likelihood from incomplete data via the EM algorithm[J]. Journal of the royalbstatistical society Series B L methodologca.l

[6]Skrbic B,Durisic-Mladenovic N.PRINCIPAL component analysis for soil contamination with organochlorine compounds[J]. Chemosphere 2007,68(11):2144-2152

[7]Lancichinetti A,Fortunato S,Radicchi F.Benchmark graphs for testing community detection algorithms[J].physical review E,2008,78(4):046110.

[8]http://www-personal.umich.edu/~mejn/netdata/.