DOMINANT EIGENVECTOR AND EIGENVALUE ALGORITHM IN SPARSE NETWORK SPECTRAL CLUSTERING

B.J. YANG

Taiyuan University, Shanxi, Taiyuan,030012, China.
ClarissaPfg@yahoo.com

***Abstract*— The sparse network spectrum clustering problem is studied in this paper. It tries to analyze and improve the sparse network spectrum clustering algorithm from the main feature pair algorithm. The main feature pair algorithm in the matrix calculation is combined with the spectral clustering algorithm to explore the application of the main feature pair algorithm on the network adjacency matrix. The defects of traditional main features are analyzed when the algorithm Power is used on the network of special structural features, and the advantages of the new algorithm SII algorithm is proved. The sparse network spectral clustering algorithm in this paper is based on the Score algorithm, and the main features of the algorithm are refined, analyzed and improved.**

***Keywords*— sparse network spectrum; clustering; dominant eigenvector; eigenvalue.**

I. INTRODUCTION

Community discovery is an important part of social network research. Spectral clustering is used widely for its low computational complexity and efficient recognition of nonlinear clusters (Barucca *et al.*, 2016) The Score algorithm is a spectral clustering algorithm that can be used for sparse networks with large differences in node degrees. The method of standardizing the feature space by the adjacency matrix main feature vector can achieve a more stable clustering output (Lee *et al.*, 2016a). However, the choice of its algorithm has not been studied as the dominant eigenvector of the core part of the Score algorithm. That is to say, the good output of Score is based on the premise that its feature vector is accurate and known (Lee *et al.*, 2016b). How the different algorithms of feature vector and eigenvalue affect Score will be the focus of this paper.

II. STATE OF THE ART

The spectral clustering algorithm calculates the similarity between the data points first thereby constructing a similarity matrix and calculating the eigenvalues and eigenvectors of the corresponding matrix (such as the adjacency matrix or the Laplacian matrix) unlike the general clustering method, which uses all sample data as the basis of clustering (Tavşanoğlu, 2016). Then, the appropriate feature vector is selected to construct the feature vector matrix to complete the clustering according to different segmentation criteria. It can be said that spectral clustering is an algorithm for clustering sample data based on the feature vector structure of the sample similarity matrix (Wei *et al.*, 2016).

It is especially important in the whole spectral clustering algorithm for the calculation of similar matrix eigenvectors, especially the calculation of the first few dimensional eigenvectors of the matrix. Matrix eigenvalue decomposition is one of the most fundamental and important aspects of matrix theory, and its research has been around for a long time (Mondal *et al.*, 2016). Up to now, matrix eigenvalues, eigenvectors, eigenfunctions and maps are still hot topics in the field of matrix algebra. The research on matrix eigenvalue problem mainly includes the distribution of matrix eigenvalues, spectral estimation, eigenvalue upper and lower bound estimation and approximate solution of high-order matrix eigenvalues (Rammohan *et al.*, 2016).

The matrix feature pairs refer to the matrix eigenvalues and their corresponding eigenvectors (Gusrialdi *et al.*, 2017). The second is the main feature pair algorithm, such as the most traditional algorithm, the power method, the minimum eigenvalue of the computational matrix and the inverse power method of the corresponding eigenvectors, Rayleigh quotient iteration and translation method for the acceleration of power method (Lekić, 2017).

Therefore, it is often not necessary to calculate all pairs of matrix features, and it is only necessary to give an estimate for the feature pairs of a certain number of dimensions before. In addition, the algorithm for solving the first feature pair is extended, and can also be used to calculate the first k feature pairs of the corresponding matrix iteratively. Therefore, the study of the main feature algorithm plays an important role in the spectral clustering algorithm.

III. METHODOLOGY

A. Origin Shift Method (SPI, Shifted Power Iteration)

Since the convergence speed of the power method depends mainly on the size of $\left|\frac{\lambda_2}{\lambda_1}\right|$, this value is always less than 1, and the smaller the value, the faster the convergence speed be. So, the translation method can be used, that is, apply the power method to A. If μ is properly selected, the modulus of the dominant eigenvalue of the matrix and the modulo of other eigenvalues are larger. That is, $\left|\frac{\lambda_2 - \mu}{\lambda_1 - \mu}\right| < \frac{\lambda_2}{\lambda_1}$, which serves the purpose of acceleration.

The translational inverse power method has the following convergence properties:

$$\lim_{k\to\infty} x^{(k)} = \frac{a_j}{\max(|a_j|)}$$

$$\lim_{k\to\infty} m^{(k)} = \lambda_j - \mu$$

Where λ_j is the eigenvalue closest to the matrix distance μ, and a_j is its corresponding eigenvector. It should be noted that the performance of the algorithm depends on the selection of the parameter μ. If μ is too small, the acceleration effect is not obvious; if μ is too large, the order of the size of the feature roots may be changed, so that the number of iterations does not decrease, and even the case of convergence to other feature roots occurs.

Algorithm1 SPI:

（1）Select initial iteration vector $x^{(0)} \neq 0$, 令 $m^{(0)} = \max\left(\left|x^{(0)}\right|\right)$, $y^{(0)} = \frac{x^{(0)}}{m^{(0)}}$

（2）For k=0:T，execute cycle

$$x^{(k+1)} = (A - \mu I)y^{(k)}$$

$$y^{(k+1)} = \max(|x^{(k+1)}|)$$

$$y^{(k+1)} = \frac{x^{(k+1)}}{m^{(k+1)}}$$

（3）If $\left|m^{(k+1)} - m^{(k)}\right| < \varepsilon$，Then finish the iteration.

（4）Output main feature pair $\left(m^{(k)} + \mu, \frac{x^{(k)}}{\left\|x^{(k)}\right\|_2}\right)$

B. SQI (Specific Rayleigh quotient iteration)

The Rayleigh quotient is used in the SQI algorithm as the eigenvalue approximation in each iteration, i.e. $z^{(k)}$, to translate the matrix.

It should be noted that the results of Rayleigh quotient iterations are more dependent on the choice of initial values.

Algorithm 2 SQI:

（1）chose Initial iteration variable $x^{(0)} \neq 0$, 令 $m^{(0)} = \max\left(A\omega^{(0)}\right)$, $y^{(0)} = \frac{x^{(0)}}{\left\|x^{(0)}\right\|_2}$,

（2）For k=0:T，execute the circle

Solve $(m^{(k)}I - A)x^{(k+1)} = y^{(k)}$ obtain $x^{(k+1)}$

$$y^{(k+1)} = \frac{x^{(k+1)}}{\left\|x^{(k+1)}\right\|_2}$$

$$m^{(k+1)} = y^{(k+1)} \cdot Ay^{(k+1)}$$

（3）If $\left|m^{(k+1)} - m^{(k)}\right| < \varepsilon$，Then finish the iteration

（4）Output main feature pair $\left(m^{(k)}, \frac{x^{(k)}}{\left\|x^{(k)}\right\|_2}\right)$

When the initial value $z^{(0)}$ is far from the actual dominant eigenvalue of the matrix and is closer to other eigenvalues, the convergence will fall into local optimum. That is, $z^{(k)}$ will

converge to the eigenvalue closest to the initial value, so the result obtained by the algorithm is not necessarily the main feature pair of the matrix.

C. Sll (Shifted inverse iteration)

The Sll algorithm is basically similar to the SQI algorithm, and the difference between the two is only the calculation method of the iterative term $m^{(k)}$.

The Sll algorithm no longer uses Rayleigh quotient as an iteration term, but instead makes $m^{(k)} = \max \frac{Ax^{(K)}}{x^{(K)}}$, that is, $m^{(k)}$ takes the value of the maximum value of the ratio of the elements of $Ax^{(k)}$ and $x^{(k)}$.

For the non-negative irreducible matrix A, there is the Collatz-Wielandt formula:

$$\sup_{x>0} \min_{i \in E} \frac{(Ax)_i}{x_i} = \lambda_1 = \inf_{x>0} \max_{i \in E} \frac{(Ax)_i}{x_i}$$

The new mode calculation $z^{(k)}$ is always larger than the true main eigenvalue λ in the iterative process, and monotonically decreasing to converge to λ, so there is no case of error convergence caused by changing the order of eigenvalues.

IV. RESULT ANALYSIS AND DISCUSSION

A. Experimental environment and data set introduction

The experimental data sets used in this article are all real network data sets. This group of experiments used a number of benchmark data sets, namely the Karate Club Network, the Canadian Dolphin Social Network, the 2004 US political book purchase network, the Facebook site's 0-node-centric ego network, the 2006 Network of Network Scientists, the e-mail (core) network among members of a European research institution, the 2005 American politician blog network, the Facebook website with a 107-node individual network, and the US Western Power Grid.

Each network is undirected and has no right, and the number of nodes N, the number of sides M, the network density s, and the aggregation coefficient C respectively included in each network are as shown in Table 1. The calculation method of the network density adopts the ratio of the number of edges existing actually in the network to the maximum possible number of sides, namely:

$$S = \frac{2M}{N(N-1)} \tag{2}$$

Table 1 basic information of dataset

Data set	Name	Number of nodes	Edge number	S	C
Karate club	Karate	34	78	13.90%	0.285
Canadian dolphin data set	Dolphin	62	159	8.40%	0.286
American political book purchase in 2004	Books	105	441	8.07%	0.488
American university football team	Football	115	613	9.35%	0.403
Collaboration data of network scientists in 2006	Cooper ation	246	583	1.93%	0.503
0-ego network of Facebook dataset	Oego	324	2514	4.80%	0.426
E-mail interaction between members of a European research institution (core)	Email	986	16064	3.18%	0.267
107-ego network of Facebook dataset	107ego	1034	26749	5.00%	0.505
2005 American statesman blog network	Blog	1222	16714	2.24%	0.226
Topological structure of western United States Power Network	Pewer	4941	6594	0.05%	0.034

B. Comparison of Algorithms for Computing Vectors of Feature Vectors

Experts and scholars have focused on the calculation of eigenvalues for the research and improvement of matrix main feature pair algorithms for a long time. Since the feature vector is used for clustering in spectral clustering, the dominant eigenvector is also used to normalize the feature space in the Score

algorithm. Therefore, the analysis and application of the feature pair algorithm are more suitable for spectral clustering. This section discusses the performance of different algorithms in computing feature vectors.

The previous section mentions that the power method has a problem of slow convergence when the second eigenvalue is closer to the first eigenvalue. If re-examining the above algorithm from the perspective of feature vector, the second norm normalized power method, Rayleigh quotient acceleration and another Aitken acceleration method only change the calculation method of the approximate eigenvalue at each iteration. The calculation method of the feature vector is actually the same, so the number of iterations required to reduce the convergence of the feature root reaches the acceleration feature root convergence, and the calculation accuracy of the feature vector is sacrificed.

The algorithm is improved in order to measure the convergence of the feature vector and eliminate the influence of the feature root calculation method on the number of convergences of the algorithm. It modifies the convergence condition in the algorithm from the difference $\varepsilon_k < \varepsilon_0$ between the iterative computational eigenvalues to the difference $\eta_k < \eta_0$ between the eigenvectors. The cosine similarity of the calculated eigenvectors in each iteration is selected to measure the closeness of the eigenvectors in the two iterations. Since the eigenvectors calculated by the algorithm are all standardized, it is only necessary to calculate the vector inner product, and then calculate the absolute value of the value and the difference η_k of 1.

$$\eta_k = 1 - |\cos\theta_k|$$
$$1 - \left|\frac{v_{k-1} v_k}{\|v_{k-1}\| \|v_k\|}\right|$$
$$1 - |v_{k-1} \cdot v_k| \quad (3)$$

Thus, the smaller the η_k, the closer the eigenvector estimates in each iteration and the more convergent be.

1. **Number of iterations**

As can be seen in Table 2, it is reflected the number of iterations of each data set under different eigenvector convergence conditions. It can be seen that under each convergence condition, as the ratio of the second eigenvalue to the main eigenvalue increases, the number of convergences of the power method increases, and the growth rate gradually increases. The convergence times of each data set under different convergence conditions are compared. As η_0 decreases, the stricter convergence condition, the more convergence times is required by the power method.

In Table 3 is illustrated the number of iterations of the SQI algorithm and the SII algorithm for each data set under different eigenvector convergence conditions. It can be seen that the number of iterations of the two algorithms is relatively close. Compared with the table, it is found that both algorithms have fewer iterations than the power method, regardless of whether the second eigenvalue and the first eigenvalue are close or not.

This reduction is especially noticeable when the second eigenvalue is close to the first eigenvalue, and both algorithms reduce the number of iterations on each data set to less than 10.

At the same time, unlike the power method, the number of convergences increases significantly with the strict convergence conditions.

As η_0 decreases, the number of iterations of the two algorithms remains the same or increases by one. Then, under the convergence conditions of different primary and secondary eigenvalues and different eigenvectors, the acceleration of the two new algorithms relative to the power method is obvious. In the time-consuming of the algorithm, the power method only needs to complete the basic matrix vector multiplication in each iteration, and the SQI and SII algorithms need to calculate the matrix equation, so each iteration takes a long time.

However, when the primary and secondary feature roots are close, the new algorithm still shows its advantages. For example, in the Books network, the number of convergences $\eta_k < 10^{-16}$ required to calculate the feature vector using the power method is as high as 520 times, and the number of iterations of the SQI and SII algorithms is less than 10 times. Since the number of network nodes is small, in order to reduce the error, we take the average time of 10 algorithm runs for comparison. The time required for the power method is 0.127 s, while the SII and SQI algorithms only need 0.022 s and 0.031 s.

2. **Calculation accuracy**

The following section focuses on the accuracy of the algorithm for the calculation of matrix feature vectors. We choose the principal eigenvector of the QR iterative calculation as the standard value v_0, and calculate the error ξ between the main eigenvector and the standard value calculated by each algorithm, and the calculation method is the same as (4-1).

Table 2. Iteration times of power law under convergence conditions of different eigenvectors.

Data set	$\frac{\lambda_2}{\lambda_1}$	$\eta_k < 10^{-5}$	$\eta_k < 10^{-10}$	$\eta_k < 10^{-16}$
Oego	0.4634	7	14	23
Email	0.4786	6	12	21
107ego	0.5018	8	16	26
Karate	0.7400	14	28	47
Cooperation	0.7956	18	42	72
Blog	0.8091	15	42	75
Dolphin	0.8407	18	51	93
Football	0.8607	11	49	98
Books	0.9738	46	263	520

Table 3. Iteration times of SQI algorithm and SII algorithm under different eigenvectors convergence conditions.

Data set	$\eta_k < 10^{-5}$		$\eta_k < 10^{-10}$		$\eta_k < 10^{-16}$	
	SQI	SII	SQI	SII	SQI	SII
Oego	7	5	7	6	8	7
Email	6	6	7	7	9	8
107ego	6	5	6	6	7	7
Karate	4	5	5	6	5	7
Cooperation	6	6	7	7	8	7
Blog	7	7	7	8	8	9
Dolphin	5	5	6	6	6	7
Football	3	4	4	5	4	6
Books	7	6	7	7	8	9

Table 4. The $\eta_k < 10^{-16}$ convergence time of each algorithm.

Data set	Number of nodes	$\frac{\lambda_2}{\lambda_1}$	pow	SQI	SII
Oego	324	0.4634	0.045	0.37	0.322
Email	986	0.4786	0.276	11.009	9.721
107ego	1034	0.5018	0.363	8.334	8.747
Karate	34	0.7400	0.005	0.02	0.027
Cooperation	246	0.7956	0.058	0.171	0.138
Blog	1222	0.8091	3.093	15.798	17.185
Dolphin	62	0.8407	0.006	0.009	0.888
Football	115	0.8607	0.023	0.038	0.023
Books	105	0.9738	0.127	0.022	0.031

Table 5. The $\eta_k < 10^{-5}$ Calculation of eigenvector and standard error by power method and SII algorithm.

Data set	Pow	Sll
Oego	3.42×10^{-6}	4.44×10^{-16}
Email	1.39×10^{-6}	1.73×10^{-14}
107ego	7.53×10^{-6}	8.88×10^{-16}
Karate	5.79×10^{-5}	3.33×10^{-16}
Cooperation	1.15×10^{-4}	4.44×10^{-16}
Blog	1.55×10^{-4}	$<1\times10^{-16}$
Dolphin	2.45×10^{-4}	5.88×10^{-15}
Football	1.27×10^{-4}	4.44×10^{-16}
Books	1.36×10^{-2}	8.07×10^{-13}

Table 5 compares the error between the eigenvector and the standard value calculated by the power method and the SII algorithm when $\eta_k < 10^{-5}$.

$$\xi = 1 - |\cos\delta_k|$$

$$1 - \left| \frac{v_0 \cdot v_k}{\|v_0\| \|v_k\|} \right|$$

$$1 - |v_0 \cdot v_k|$$

It can be seen that for each data set, the eigenvector error calculated by the SII algorithm is smaller than the power method. When $\eta_k < 10^{-5}$, the magnitude of the error between the feature vector calculated by the power method and the standard value is greater than 10-6, and the eigenvector error of the power method increases, and the maximum error reaches 10-2 on the Books network as the ratio of the second eigenvalue to the first eigenvalue increases. In contrast, the SN calculates the eigenvector and the standard value error is much smaller, both less than 10-13. It can be seen that the accuracy comparison of the eigenvectors calculated by the power method depends on the convergence condition, while the SII algorithm is not. When the convergence condition is lost, the convergence accuracy can be achieved.Then we analyze the convergence trend of the algorithm in each iteration. It is found that when both the power method and the SII algorithm select the unit vector as the initial iteration vector, the same phenomenon occurs on each data set. That is, in the initial several (generally 2~3) iterations, the eigenvectors calculated by the power method are closer to the standard value than the SII algorithm. The proximity is overridden by the SII algorithm after several iterations. The SII algorithm approximates the exact value at a faster convergence rate, and the accuracy is higher than the power method under the same number of iterations.

V. CONCLUSIONS

The new algorithms SQI and SII algorithm get rid of the dependence of the convergence speed of the algorithm on the proximity of the primary and secondary eigenvalues, and reduce the number of convergences greatly. In the experiment, the convergence is completed with less than 10 iterations on each data set, and the calculation of accuracy of the eigenvalue and eigenvector is also improved greatly. There is only a convergence error of 10-16 under the convergence of $\varepsilon_0=10^{-6}$. Between the two, the number of iterations is quite the same, often one or two iterations. The convergence effect of the SQI algorithm is unstable, and error convergence occurs in a matrix with a higher dimension

REFERENCES

Barucca, P., D. Tantari and F. Lillo, "Centrality metrics and localization in core-periphery networks", *Journal of Statistical Mechanics: Theory and Experiment*, **2016(2)**, 023401 (2016).

Gusrialdi, A. and Z. Qu, "Distributed Estimation of All the Eigenvalues and Eigenvectors of Matrices Associated with Strongly Connected Digraphs", *IEEE Control Systems Letters*, **1(2)**, 328-333 (2017).

Lee, A.B. and R. Izbicki, "A Spectral series approach to High-Dimensional Nonparametric Regression", *Electronic Journal of Statistics*, **10(1)**, 423-463 (2016).

Lee, J.O. and K. Schnelli, "Extremal eigenvalues and eigenvectors of deformed Wigner matrices", *Probability: Theory & Related Fields*, **164(1-2)**, 165-241 (2016).

Lekić, A., "Simulation of PWM DC-DC converters using eigenvalues and eigenvectors", *Journal of Electrical Engineering*, **68 (1)**, 13-22 (2017).

Mondal, S, and M. Pal, "Similarity relations, eigenvalues and eigenvectors of bipolar fuzzy matrix", *Journal of Intelligent and Fuzzy Systems*, **30(4)**, 2297-2307(2016).

Rammohan, L.R. and M.A. Dayananda, "Ternary diffusion path in terms of eigenvalues and eigenvectors", *Philosophical Magazine*, **96(10)**, 938-954 (2016).

Tavşanoğlu, V., "Decomposition of the Nodal Conductance Matrix of a Planar Resistive Grid and Derivation of Its Eigenvalues and Eigenvectors Using the Kronecker Product and Sum with Application to CNN Image Filters", *IEEE Transactions on Circuits & Systems I: Regular Papers*, **63(12)**, 2169-2179 (2016).

Wei, L. and L. Bing, "Combining eigenvalues and variation of eigenvectors for order determination", *Biometrika*, **103(4)**, 875-887 (2016).

Received: December 15th 2017
Accepted: June 30th 2018
Recommended by Guest Editor
Juan Luis García Guirao