Key Node Identification Method for Aviation Network Based on Support Vector Machines

Haiqing Huang[1,a], Xusheng Gan[2,b]*, Xu Yao[2,c], Fei Liu[2,d]

[1]XiJing College, Xi'an, Shaanxi, 710123, China

[2]Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an, Shaanxi, 710051, China

[a]ganxusheng123@163.com, [b]gxsh15934896556@qq.com, [c]yaoxu11@163.com, [d]liufei1987@126.com

Keywords: Aviation Network, Support Vector Machines, Key Node Identification, Analytic Hierarchy Process

Abstract: Accurately identifying the key nodes of aviation network through technical means has important theoretical significance and reference value for the normal operation of aviation network in peacetime and defense and repair in wartime. A key node identification method based on Support Vector Machines (SVM) is proposed. Firstly, it evaluates the comprehensive importance of nodes based on analytic hierarchy process (AHP). Then, it selects three simple indices and establishes the importance evaluation model based on the mapping relationship between simple indices and comprehensive importance of SVM. The simulation on American aviation network indicates that it can obtain satisfactory identification effect.

1. Simple index of nodes

For aviation network, the research on key airports is mainly index analysis. When the index is small, the evaluation accuracy is not ideal, and when the index is more comprehensive, the calculation time complexity is high. The simple index value is the training knowledge database. In this paper, node degree value, point strength, and K-shell value are selected as simple indices.

Node degree value: reflects the number of connections between a single node and adjacent nodes in the network. it can be defined as the number of direct connections between the nodes

$$k_i = \sum_j a_{ij} \tag{1}$$

where a_{ij} is the connection status between two nodes. If the node v_i and v_j does not have a direct connection, then $a_{ij} = 0$, otherwise, $a_{ij} = 1$.

Point strength: mainly refers to the edge weights in the aviation network, that is, the route flow. The point strength S_i can be expressed as

$$S_i = \sum_{j \in N_i} w_{ij} \tag{2}$$

where w_{ij} denotes the weight of the edge directly connected to the node v_i, and N_i represents the set of adjacent nodes of the node v_i. The closer the connection between the surrounding airport and the airport node is, the greater the connection edge weight.

K-shell value: as shown in Fig. 1, K-shell method is a representative algorithm for sorting several nodes. According to the degree of nodes or other indices, the nodes of the network shell are stripped layer by layer. The later the stripping, the more important the node [1]. The specific steps: search for nodes with degree 1 in the network, and delete such nodes and their connection edges. After deleting these nodes, the network structure changes, delete nodes with degree 1 and their edges, and follow this process, continue to delete nodes until no nodes with degree 1 are included in the network. The

  DOI: 10.25236/cseem.2020.003

shell composed of the deleted nodes is regarded as a 1-shell (ie. *Ks*=1). In the same way, continue to remove nodes with node degree 2 as 2-shells, and so on, until all nodes are deleted. This method performs coarse-grained sorting on nodes. Although the accuracy is not high, it reflects the global nature of nodes.

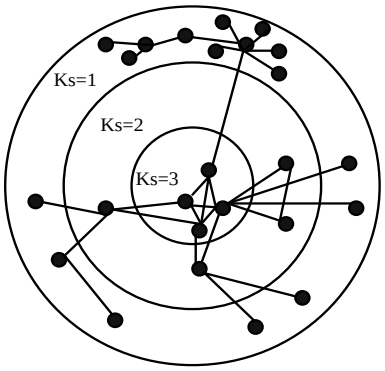


Fig. 1 Schematic diagram of K-shell method

For the node v_i, its K-shell value is Ks_i, and the larger the value, the more important the node.

2. Complexity index of nodes

In the field of complex networks, methods to identify key nodes mainly include social network analysis methods and systems scientific analysis methods.

For social network analysis methods, closeness centrality and betweenness centrality can be selected as evaluation indices.

Closeness centrality (*CC*): Calculate the average distance between the node v_i and the remaining nodes in the network to solve the problem of special values. If the distance between v_i and other nodes is smaller than that between v_j and other nodes, it can be considered that CC of v_i be larger than that of v_j. Generally, the node closest to the center has the best view of the information flow. Suppose the network contains *n* nodes, then the average value of the shortest distance from v_i to the remaining nodes in the network is

$$d_i = \frac{1}{n-1}\sum_{j\neq i} d_{ij} \tag{3}$$

If d_i is smaller, it means that v_i is closer to the remaining nodes of the network, and the reciprocal of d_i can be defined as *CC* of the node v_i

$$CC(i) = \frac{1}{d} = \frac{n-1}{\sum_{j\neq i} d_{ij}} \tag{4}$$

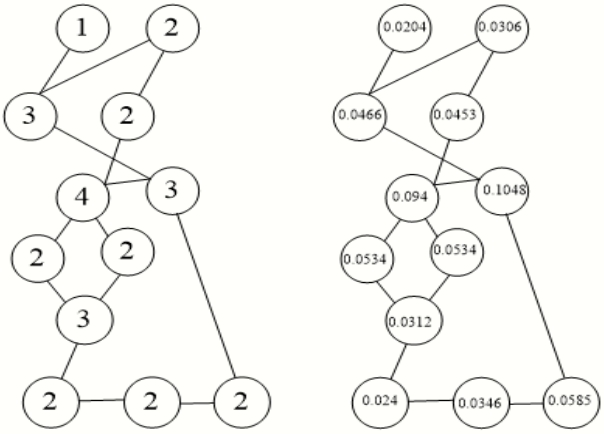
In the above formula, the larger $CC(i)$ is, the closer v_i is to the center of the network, the more important the location and the greater the importance. Fig. 2 is a comparison of the effect between node degree method and closeness degree method. It can be seen that closeness degree can distinguish the importance of nodes more accurately than node degree.

Betweenness centrality (*BC*): reflects the centrality of nodes in the overall network. The betweenness $BC(k)$ of the node v_k refers to the proportion of the number of shortest paths between all pairs of nodes in the network through the node v_k to the total number of shortest paths

$$BC(k) = \sum_{i \neq j} \frac{\sigma_{ij}(k)}{\sigma_{ij}} \tag{5}$$

where $\sigma_{ij}(k)$ is the number of shortest paths through v_k between v_i and v_j, and σ_{ij} is the number of shortest paths between v_i and v_j.

For system scientific analysis methods, node deletion method can be used. The idea of node deletion method: calculate the network performance after deleting a node and compare it with the original network. The greater the change in network performance, the more important the node. The network performance can be measured by network connection density and network efficiency.



Node degree Closeness centrality

Fig. 2 Comparison of the effects of two methods

Network connection density (*LD*): In an unweighted network, the connection density refers to the ratio of the existing connection edge to the possible connection edge in the network [2]. For aviation networks, a weighted connection density can be defined as

$$LD = \frac{\sum_{i}^{n} \sum_{j}^{n} a_{ij} w_{ij}}{2n} \tag{6}$$

where n is the total number of current network nodes. If v_i and v_j are directly connected, $a_{ij} = 1$; otherwise, $a_{ij} = 0$, w_{ij} is the weight of the connection edge between nodes. The larger *LD*, the higher the overall heterogeneity, the larger the network traffic flow, and the better the overall network performance.

Network efficiency (*NE*) is the average value of the sum of reciprocal of the distances between all nodes

$$NE = \frac{1}{N(N-1)} \sum_{i \neq j} 1/d_{ij} \tag{7}$$

where N is the total number of nodes in the network. *NE* reflects the difficulty of network information transmission. The larger *NE*, the smoother the information transmission and the stronger the survivability.

3. Support Vector Machines (SVM)

SVM is a machine learning technique developed in the mid-1990s. It introduced the concept of structural risk and adopted the idea of kernel mapping [3][4], which overcomes the harsh demands of neural network method on sample size and avoids Learning, local minima and dimensions of disaster

and other issues.

For the nonlinear regression, suppose that the training set is $\{(x_1, y_1),\cdots,(x_n, y_n)\}$, $x_i \in R^d$. Firstly, the data x in the training set is mapped into a high-dimensional feature space by a nonlinear function $\phi(\cdot)$, and then in the feature space the linear regression is carried out, which can obtain the effect of nonlinear regression in the original input space. So the nonlinear regression problem can be described as looking for a nonlinear function $f \in F$ to minimize the empirical risk, namely use

$$f(x) = (w \cdot \phi(x)) + b \tag{8}$$

to fit the sample data and ensure a good generalization ability. $w \in R^d$, $b \in R$. Then the regression estimate problem is expressed as

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{9}$$

$$s.t. \begin{cases} y_i - (w \cdot \phi(x_i)) - b \le \varepsilon + \xi_i \\ (w \cdot \phi(x_i)) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{10}$$

where the penalty parameter ($C > 0$) can maintain the balance between the flatness of regression function f and the number of sample points which the deviations are more than ε. The Formula (9) and (10) can be transformed into the following dual optimization problem

$$\min \ \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \sum_{i=1}^{n}\varepsilon(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{11}$$

$$s.t. \ \begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \tag{12}$$

where $K(x_i, x_j) = (\phi(x_i) \cdot \phi(x_j))$ is called the kernel function. Then the regression estimate function can be expressed as

$$f(x) = \sum_{x_i \in \mathrm{SV}} (\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{13}$$

where SV is the support vector. It can be seen that the estimation of regression function only needs to calculate the kernel function without using the nonlinear mapping function $\phi(\cdot)$. In the paper, Least Square Algorithm is selected as the learning algorithm of SVM [5].

4. Evaluation process for node importance

1) Construct training samples. Some nodes are randomly generated from the network, and the simple node index value X such as degree value, point strength and K-shell value are calculated. At the same time, the complex index values such as closeness centrality, betweenness centrality, network connection density and network efficiency are determined.

2) Determine the comprehensive importance. Obtain the weight of each complex index by Analytic Hierarchy Process (AHP) [6]: W_{BC}=0.5789; W_{NE} = 0.2055; W_{CC}=0.1592; W_{LD}= 0.0565, and then calculates the comprehensive importance Y =[Y_1, Y_2,$\cdots$,Y_n]$^{\mathrm{T}}$ of the above random nodes by Y_i=0.5789BC_i+0.2055NE_i+0.1592CC_i+0.0565LD_i.

3) Train the evaluation model. Taking X as input and Y as output, train the SVM evaluation model to learn its intrinsic relationship.

4) Perform the evaluation process. For the new nodes in the network except for the training samples, only the simple indices X_t of the new nodes need to be calculated and input into the SVM evaluation model to calculate the comprehensive importance Y_t of the new nodes, and then complete the identification of key nodes.

5. Simulation

Consider a random network $G=\{V, E, W\}$, which contains 600 nodes and 6000 edges. The purpose of the experiment is to test the effectiveness of the SVM key node identification method, that is, judge whether SVM can accurately learn the relationship between simple indices and comprehensive importance.

According to the steps of key node identification algorithm, 60 nodes are randomly selected as the training samples of SVM. The weights of the complex index *CC*, *BC*, *LD*, and *NE* are obtained through AHP as 0.1592, 0.5789, 0.0565, and 0.2055 respectively. The weighted sum of complex indices is calculated as the comprehensive Importance value Y, and then calculate the corresponding simple index value X. Before SVM training, the cross-validation method is used to automatically search for the best parameters c and σ in $\log_2 c \in [-10,10]$ and $\log_2 \sigma \in [-10,10]$. The parameter optimization process is shown in Fig. 3.

In the figure, Root Mean Square Error (RMSE) of the actual values and predicted values for the importance of complex indices is given

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(Y - Y_i)^2} \tag{14}$$

It can be seen that the bottom of the grid graph is relatively flat, indicating that the RMSE can be minimized by two parameters in a large range of values. Therefore, it is easier to find the best parameters c and σ.

After training, 60 nodes other than the training samples are randomly selected as test nodes, and calculate their simple index X_t and input the KELM model to obtain the test result Y_t. Comparison of test results and actual values is shown in Fig. 4.

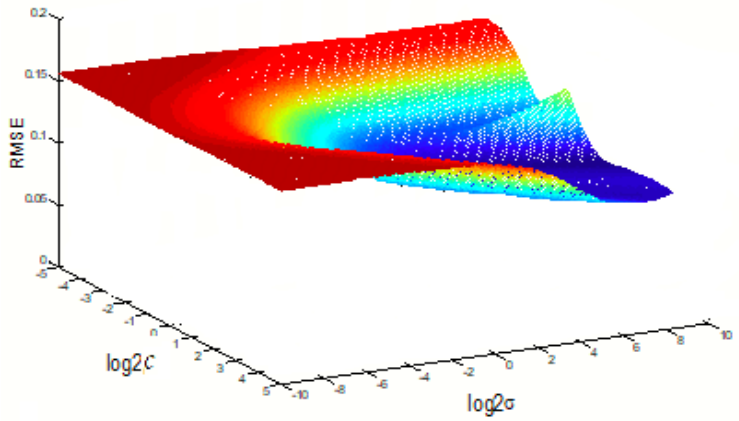


Fig. 3 Parameter optimization process for random network

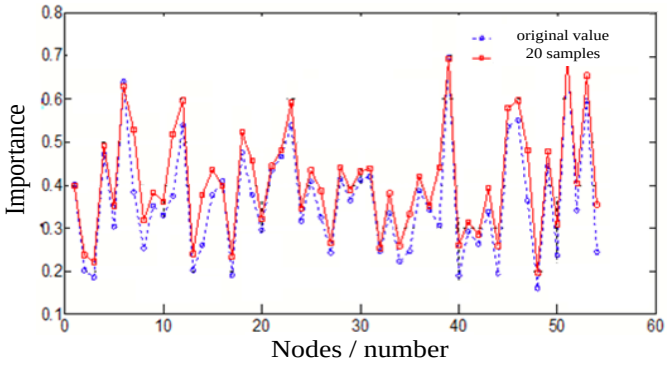


Fig. 4 Comparison of test results and actual values

It can be seen that the output result of SVM is very close to the original importance value, indicating that the method in this paper is accurate and feasible. It can be known from Table 2 that the time complexity required to evaluate the importance of nodes using complex indices is $O(N^3)$, while SVM evaluation only requires $O(N'^2)$, where N' is the number of training samples. Therefore, with this method, the comprehensive importance of nodes can be quickly obtained through simple and Less time-consuming indices.

6. Conclusions

Use SVM to learn the mapping relationship between comprehensive importance values and simple indices. For most of the remaining nodes, only simple indices are needed to obtain their comprehensive importance and node ranking. It solves the traditional single index problem without considering the edge weights. This improves the accuracy of node sorting, reduces the computational complexity, and saves a lot of time. Examples verify its effectiveness and feasibility.

References

[1] Z. H. Liu, C. Jiang, J. Y. Wang, et al. The node importance in actual complex networks based on a multi-attribute ranking method. Knowledge- Based Systems, 84, (2015), 56-66

[2] H. Y. Ying, A. Jaiswal, T. Hollstein, et al. Deadlock-free generic routing algorithms for 3-dimension networks-on-chip with reduced vertical link density topologies. Journal of Systems Architecture, 59(7), (2013), 528-542

[3] V. Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag, (1999)

[4] V. Vapnik. Statistical Learning Theory. New York: Wiley, (1998)

[5] J. A. K. Suykens, V. T. Gestel, J. De Brabanter, et al. Least squares support vector machines. World Scientific, (2002)

[6] G. H. Yu, Y. J. Zhang, H. J. Huang, et al. A new method for consistency test of judgment matrix in AHP. Mathematics in Practice and Theory, 47(22), (2017), 189-198