Study on Network Complexity Based on Clustering Algorithm

Xinlei Li*

Henan Normal University, Henan Xinxiang 453007, China

Abstract: By introducing concept of complex system into computer network system, the computer network can be regarded as a complex system with huge volume of data and information contents, including local area network (LAN), Internet and wide area network (WAN), etc. The relationships of peak and valley of network and the conflict and competition are found out through tracking network. then the relationship among each member is optimized so as to achieve the reasonable operation of complex system. In the case of classifying the researches on network complexity, the computer system is divided into TCP/IP system, ISO/OSI reference model, and so on, or it is classified in accordance with competitive characteristics of different networks. Based on clustering algorithm, this study presents the network complexity in terms of network flow, and compares the accuracy of different clustering algorithms.

Keywords: Network complexity, Clustering algorithm, EM clustering algorithm.

1. INTRODUCTION

The computer network has a vast number of users which had reached 791,000,000 based on the statistics in 2005, while the growth rate is up to 7% [1-4]. As shown in Fig. (**1**), the passageway network bandwidth in China increased annually from 2004 to 2011. What's more, the user number will grow faster with the continuous development and progress of technologies including router, switch, etc.

Domestic scholars consider that computer network is an opening but complicated system containing WAN, internet, etc. With the increase of user number and information contents, the computer LAN system is treated as a complex system in which multiple LANs superpose into a WAN, while multiple WANs form the internet [5-8]. For current computer network, the studies on its complexity mainly include the similarity of internet traffic, virus prevention and monitoring of internet, and evolution of internet structure.

2. ESTABLISHMENT OF CLUSTERING ALGORITHM MODEL

2.1. EM Clustering Algorithm

The EM clustering algorithm is defined as a clustering method which refers the maximum likelihood estimation of parameters [8-12]. Due to the imperfection of data, the maximum likelihood estimation is used to estimate the incomplete data. Therefore, EM cluster algorithm is very suitable for the condition of incomplete data.

Suppose a set

$$Z = (X, Y) \tag{1}$$

Where X denotes the observed data, and Y denotes the unobserved data, so $Z = (X, Y)$ contains the complete data, but X and Y are incomplete data.

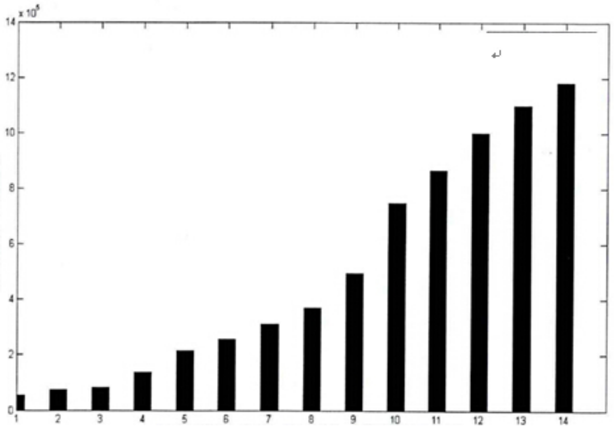

Fig. (1). Passageway network bandwidth trend in china from 2004 to 2011.

Suppose the probability density of is

$$p(X, Y / \Theta) \tag{2}$$

Θ assigned as the estimated parameter. By estimating Θ 's maximum likelihood estimation to get Z's logarithm likelihood function $L(X, \Theta)$, the estimation progress is shown as the follows:

$$L(X, \Theta) = \log \mathrm{p(X,Y \backslash \Theta)} = \int \log \mathrm{p(X,Y \backslash \Theta)} dY \tag{3}$$

EM clustering algorithm contains two steps.

Table 1. Three competition grades of inter-domain grade, LAN grade, router grade (gateway grade).

Competition Grade	Inter-domain Grade (Inter-AS Grade)	LAN Grade	Router Grade (Gateway Grade)
Network of Competition	Inter-domain	Inner LAN	Domain
Competitive Resources	Cache, Bandwidth, Processing Ability of Router	Channel,(shared device in network, like printer)	Cache, Bandwidth, Processing Ability of Router
Protocol	BGP	IEEE802 Protocol Standard	ICM, UDP, TCP,etc.
Competitor	Routing node	User Nodes	User Nodes or Routing Node

Table 2. Distinguish marks of network flow.

Mark Name	Byte Number	Meaning
Des IP	4 bytes	Server IP address
Source IP	4 bytes	Client IP address
Tcp	1 bytes	Transport layer Protocols
Des port	2 bytes	Server port number
Source port	2 bytes	Contains client port number

The logarithm likelihood function of maximized incomplete data is as below:

$$Lc(X,\Theta)=\log p(X,Y\backslash\Theta) \quad (4)$$

If Θ's estimated value after t time's iterations is $\Theta(t)$, the logarithm likelihood function of complete data after $t+1$ time's iterations is defined as below:

$$Q(\Theta\backslash\Theta(t))=E\{Lc(\Theta;Z)\backslash X;\Theta(t)\} \quad (5)$$

The newly defined by the maximization of is got.

2.2. K-means Clustering Algorithm

The algorithmic processes of K-means clustering algorithm is shown as the follows [13, 14]:

(1) Select randomly M points as clustering centers;

(2) According to nearby principle, shut the objects which are nearest these points. Thus there are clusters formed by different clustering centers;

(3) By means of iteration, change the clustering centers to get the optimal results.

Based on K-means clustering algorithm, it is thought that the closer to the clustering center, the higher the similarity is, the clusters are obtained through clustering centers.

2.3. Clustering Based on Grid Density

The clustering based on grid density is regarded as a clustering algorithm based on different grid densities. Its algorithm principle is as below:

(1) find out the convex clusters with prominent densities;

(2) Filter the low-density clusters so as to get high-density ones.

According to clustering algorithm, the similar objects are classified and closure, while the complex objects is reasonably distributed and arranged. Each object after being shut can composite a set, and these sets have similar characteristics.

If the system is classified in terms of hierarchy characteristic of system institution, computer system is divided into TCP/IP system, communication standard ISO/OSI reference model, network topology structure and distribution system of Internet domain name or it can be classified based on competitive characteristics of different network, as shown in Table **1**.

2.4. Network Flow

The data in network flowing through certain routers compose the clusters with different clustering in accordance with some forms. Then the data in each cluster form data flows on basis of time sequence, thus these data flows become network flow. In this study, the network is distinguished in terms of destination IP address, destination port, source port, source IP address and transport layer, as shown in Table **2**.

If computer 2 represents the client while computer is server, the server can receive the information from computer 2, so it can be called as network topology relationship, as shown in Fig. (**2**)

It needs specific server to complete the network communication when network connection is established between

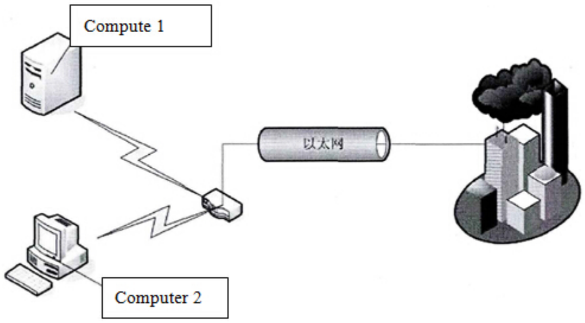

Fig. (2). Network topology relationship.

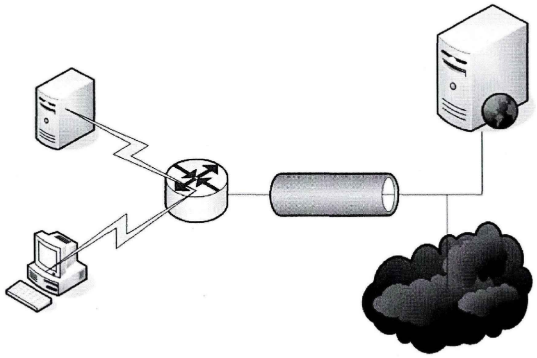

Fig. (3). Server of networking in the LAN of VNC and SSH.

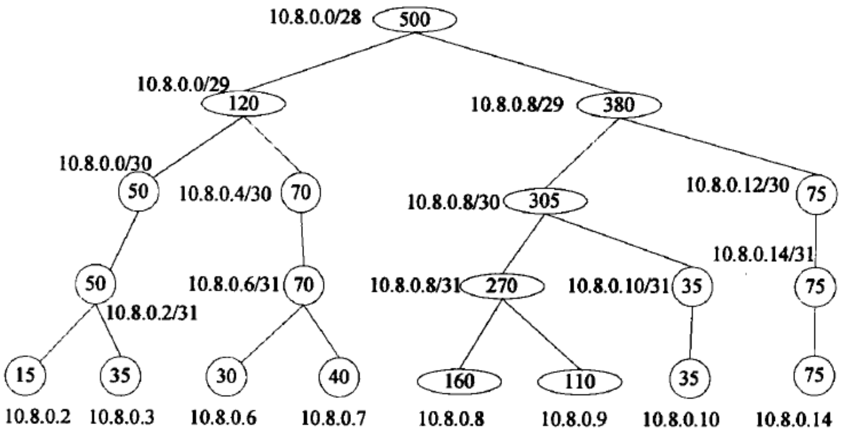

Fig. (4). Topological relation.

VNC and SSH. In LAN, the specific computer is needed to serve the network operation of VNC and SSH. As shown in Fig. (**3**), the terminal PC is considered to be the server of networking in the LAN of VNC and SSH.

Based on Fig. (**2**) and Fig. (**3**), there is the topological relation figure---Fig. (**4**).

2.5. Algorithm Flow of Clustering Algorithm in Network Flow

The recognition process of clustering algorithm to network flow is as below:

(1) Obtain the network flow in network at certain time, and then arrange these data;

(2) Formulate the division principle of network flow; and divide the data of intercepted network flow on basis of the division principle;

(3) Extract data and match them;

(4) Analyzing the network flow to determine whether the network flow can be matched with application protocol of network;

(5) Otherwise, re-division.

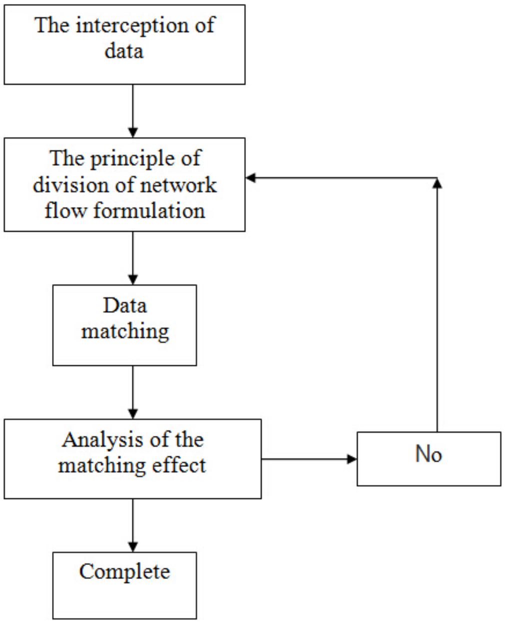
The interception of data
The principle of division of network flow formulation
Data matching
Analysis of the matching effect
No
Complete

Fig. (5). Recognition process figure of clustering algorithm to network flow.

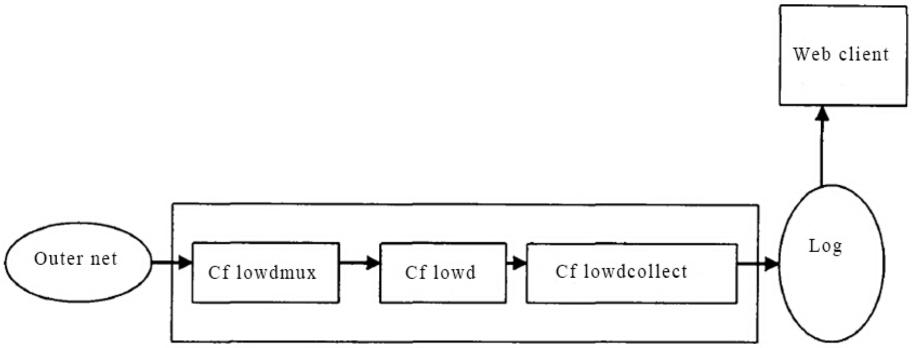

Fig. (6). Framework of C flowd's tool set.

Table 3. Comparisons of major firms' network flow technology.

Flow Name	Representative Firm	Main Version	Introduction
NetFlow	Cisco	V1,V2, V3, V4, V5	Wide application
Cflowd	JuniPer	V8, V5	Firm's following-up strength is weak
sflow	HP, NEC, A leate, Extreme, Foundry	V4, V5	Need all web's coverage
IPFIX	IETF standard specification	RFC3917	NetFlow V9
Netstream	Quidway	V8, V5, V9	NetFlow

The progress is shown as in Fig. (**5**):

During the analysis process of clustering algorithm to network flow, the Cflowd tool set is firstly used, for internet provides the NetFlow tool Cflowd. The framework is shown as in Fig. (**6**).

Presently some major firms' network technologies are shown as in Table **3**, containing tool Cflowd.

Fig. (**7**) presents the change curves of accuracy calculated by K-means clustering algorithm and EM clustering algo-

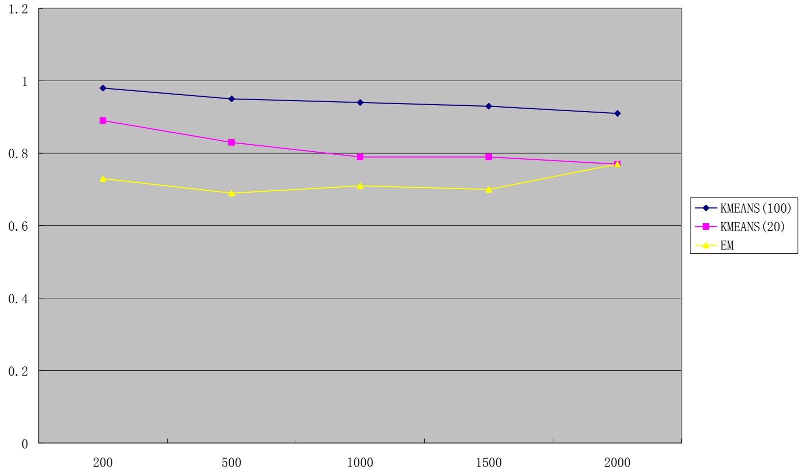

Fig. (7). Effect curves of EM and K-means algorithms with different data sizes.

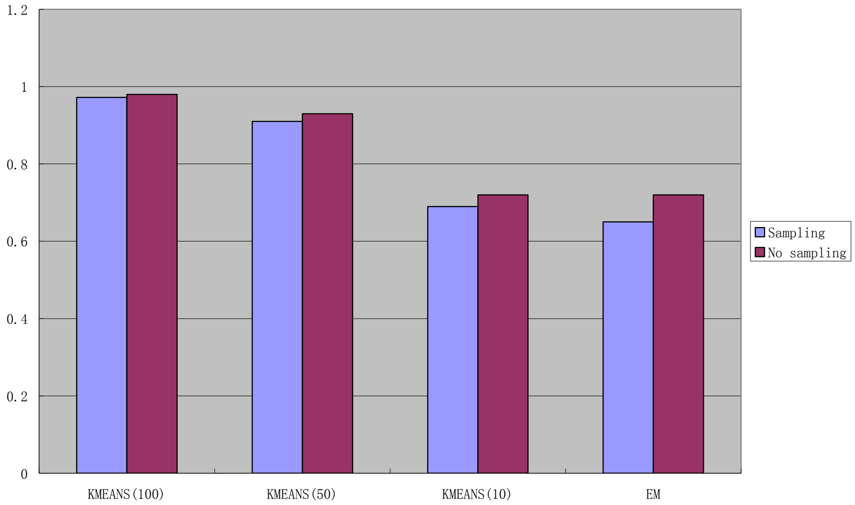

Fig. (8). Sampling accuracy is different with that of non-sampling one.

rithm. The abscissa denotes the growth of network while the ordinate represents the accuracy.

As seen in Fig. (**7**), with the increase of network flow, the accuracy of EM clustering algorithm is improved, on the contrary the accuracy of K-means clustering algorithm decreases.

Fig. (**8**) shows that sampling accuracy is different with that of non-sampling one. The sampling algorithm changes the behavior of network, which creates deviation, naturally the accuracy of clustering algorithm decreases.

CONCLUSION

By applying the concept of complex system into computer network system, the computer network can be defined as a complex system with huge volume of data and information contents, including local area network, Internet and wide area network, etc. The relationships of the peak and valley of network, and the conflict and competition are found out by means of tracking network. The relationship among each member is optimized to achieve the reasonable operation of complex system.

It is concluded that with the increase of network flow, the accuracy of EM clustering algorithm increases, while the accuracy of K-means clustering algorithm decreases. Additionally, the accuracy of sampling accuracy is different with that of non-sampling one. The sampling algorithm reduces the accuracy of clustering algorithm.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money", *Proceedings of the 27th Annual Computer Security Applications Conference*, ACM, Orlando, USA, 2011.

[2] H. Kwak, C. Lee, H. Park, S. Moon, "What is Twitter, a social network or a news media?", *Proceedings of the 19th International Conference on World Wide Web*. ACM, Raleigh, USA, 2010.

[3] W. Enck, D. Gilbert, B.-G. Chun, L.P. Cox, J. Jung, P. McDaniel, and A.N. Sheth, "TaintDroid: an information flow tracking system for real-time privacy monitoring on smartphones", *Communications of the ACM*, vol. 57, no. 3, pp.99-106, 2014.

[4] K. Kumar, and Y.H. Lu., "Cloud computing for mobile users: can offloading computation save energy?", *Computer*, vol. 43, no. 4, pp.51-56, 2010.

[5] H. Xu, W. Pu, and H. Lan, "An improved genetic algorithm for solving simulation optimization problems", *International Journal of Physical Sciences,* vol. 6, no. 10, pp. 2399-2404, 2010.

[6] J. Huang, *Adaptive Media Transport Management for Continuous Media Stream Over LAN/WAN Environment*, U.S. Patent, US 7984179, no. 7, pp. 984,179, 2011.

[7] B. Zhuge, M. Yao, L. Wan, W. Wang, and J. Lan, "Research on the reconfigurable network system based on the task decomposition", *Information Technology Journal,* vol.12, no.10, 2013.

[8] R. Kottomtharayil, P. Gokhle, A. Prahlad, M.K. Vihayan, M. K. Vijayan, D. Ngoand, V. Devassy, "*Systems and Methods for Performing Storage Operations in a Computer Network*", U.S. Patent Application, vol. 14, no. 183, p. 777, 2014.

[9] M. Meila, and D. Heckerman, "An experimental comparison of several clustering and initialization methods", *arXiv preprint arXiv,* 1301. 7401, 2013.

[10] D. A. Earl, "Structure harvester: a website and program for visualizing structure output and implementing the Evanno method", *Conservation genetics resources,* vol. 4, no. 2, pp. 359-361, 2012.

[11] T. Denœux, "Maximum likelihood estimation from fuzzy data using the EM algorithm," *Fuzzy Sets and Systems,* vol. 183, no. 1, pp. 72-91, 2011.

[12] H. Cao, H.W. Deng, and Y.P. Wang, "Segmentation of M-FISH images for improved classification of chromosomes with an adaptive fuzzy c-means clustering algorithm", *IEEE Transactions on Fuzzy Systems,* vol. 20, no. 1, pp. 1-8, 2012.

[13] S. Na, X. Liu and Y. Guan, "Research on k-means clustering algorithm: An improved k-means clustering algorithm', *Intelligent Information Technology and Security Informatics (IITSI)*, *3rd International Symposium on IEEE*, 2010.

[14] M. E. Celebi, A. K. Hassan, and A. V. Patricio, "A comparative study of efficient initialization methods for the k-means clustering algorithm", *Expert Systems with Applications,* vol. 40, no. 1, pp. 200-210, 2010

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014