A Subspace Multi-Relation Model for the Expression of System Complexity

Dongyun Yi[1,2], Yangyang Liu[2], Xiaojie Wang[2], Zheng Xie[2], Xianggui Qu[3], Chengli Zhao[2] and Chengping Hou[2]

[1] *National Key Laboratory of high performance computing, Changsha, Hunan, 410073, China;*

[2] *Department of Mathematics and System Science, National University of Defense Technology, Changsha, Hunan, 410073, China;*

[3] *Department of Mathematics and Statistics, Oakland University,Rochester, Michigan, 48309, USA.*

Abstract

Modeling expression of complexity is one of the frontier research fields in system science. Classical graphical models characterize two-relations but lack of the ability of depicting multi-relations among objects in a system. In this paper, a multi-relation expression model is proposed based the subspaces of the attribute space of objects in a system. Properties of the multi-relation model are analyzed and an algorithm is provided to find the multi-relation subspaces in a data set. The relationship between the multi-relation model and a hyper graphical model is also discussed.

Keywords Complexity; Relation; Subspace; Hyper-graph.

1 Introduction

There are two types of research in studying the complexity of a system [1-3], One considers how to characterize and compute the overall complexity of a system [4]. Boltzmann entropy in statistical physics is such a representative which gives an overall measure for the system status. The other type of research describes the interactive relationships among various objects in a system. Graphical models from graph theory or complex network are typical representatives [5]. A graphical model regards the objects in a system as vertices and relationships among objects as edges that can be undirected, directed or weighted. Graphical models have a lot of applications in the analysis of e-mail data, communication data, web data, biological data, financial data, social networks, actor networks, etc. [6-8]. However, graphical models only depict two-relation between objects, lack of the description of multi-relation in a system in practice. Recently, some research articles discuss the multi-relation using hyper-graph models [9-13], but no rigorous definitions and generating methods of hyper-graphs from data are given.

In this paper, we propose a subspace multi-relation model to characterize the

This research is supported by The Natural Science Foundation of China(No.61473302), The Basic Research Project of NUDT and The Open Fund from HPCL.

different types of relationships in the attribute space of objects. In this model, each relationship exists in its own subspace, which is a part of the attribute space. The traditional graph model is a special case of our model, which can be called one-dimension relation model.Based on the subspace clustering principle [14-16], we give the algorithm to find subspace multi-relation from the object attribute space, which is helpful to understand the complexity of data from multiple perspectives. Real data are used to verify the validity of our model. We discuss the vector hyper-graph models at the end.

2 Methods

2.1 Definitions of Subspace Multi-Relation

In this section, we define the subspace multi-relation model.

Definition 2.1 *If m objects have a common property of P, we call the m objects as m-relation.*

In the examples provided by Tables 1 and 2, there are three events and three individuals. Property P means that there are individuals appearing in the same event. In Table 1, any two individuals appear simultaneously in one event, and therefore there exists a two-relation between any two individuals, which can be expressed by the fully connected network model. In Table 2, we notice that three individuals simultaneously appear in the event A. Individuals 1, 3 appear in the event B and individuals 2, 3 appear in the event C. Thus there are one three-relation and two two-relations by Definition 2.1. The difference between Table 2 and Table 1 shows that: Firstly, three-relation cannot be obtained from two-relation; Secondly, the definition of the multi-relation needs consider the subspace arising relation. Thus we should make some modification to Definition 2.1.

Table 1 Any two individuals appear simultaneously in one event, and hence, two-relations between any two of them exist.

	Event A	Event B	Event C
Individual 1	1	1	0
Individual 2	1	0	1
Individual 3	0	1	1

Table 2 There are three individuals appearing in the event A at the same time, which indicates that a three-relation exists.

	Event A	Event B	Event C
Individual 1	1	1	0
Individual 2	1	0	1
Individual 3	1	1	1

Establishing relations among objects needs to consider the object attribute space. For example, a person can be described by an attribute space including gender, age, education, hobby, strength, event, and so on; an email attribute space includes sender, recipient, delivery time, subject, body, attachments, and people names that appear in the mail, and so on. Hence each object can be expressed by a numerical vector in its attribute space. Assuming there are m objects denoted by $x_i, i = 1, 2, \cdots, m$, of which attribute space is n-dimensional. Then each object can be represented as a n-dimensional vector $x_i = (x_{i1}, \cdots, x_{in})^T \in \mathbb{R}^n, i =$

$1, 2, \cdots, m$. In order to be able to obtain the multi-relations among objects from the attribute space, we give the definition of subspace multi-relation.

Definition 2.2 *Suppose that* $\mathbb{R}^n = X_1 \times X_2 \times \cdots \times X_n$,*and if there exists a sub-sequence denoted by* $i_1 < i_2 < \cdots < i_k, A_{i_j} \subset X_{i_j}$ *is a subset of* $X_{i_j}, j = 1, 2, \cdots, k$,*which makes the subset* $B = A_{i_1} \times A_{i_2} \times \cdots \times A_{i_k}$ *in the k-dimension subspace* $X_{i_1} \times X_{i_2} \times \cdots \times X_{i_k}$ *has a common property denoted by* P. *Then we call set B as the k-dimension subspace m-relation, where m is number of element in set B.*

Definition 2.2 shows that the relation not only means that the objects in set B have a common property, but also tells us that the relation occurs in which attribute space. As defined, the multi-relation in Table 1 can be written as $(x_1, x_2; A), (x_1, x_3; B), (x_2, x_3; C)$, and the multi-relation in Table 2 can be written as $(x_1, x_2, x_3; A), (x_1, x_3; B), (x_2, x_3; C)$. Therefore there are three one-dimension subspace two-relations in Table 1, and two one-dimension subspace two-relations and one one-dimension subspace three-relation in Table 2. We have defined the k-dimension subspace multi-relation. How can we compute subspace multi-relations in real data? The following Definition 2.3 provides a solution.

Definition 2.3 *Given* $\epsilon > 0$, *let* $\mathbb{R}^n = X_1 \times X_2 \times \cdots \times X_n$, *if there is a sub-sequence of* $1, 2, \cdots, n$ *denoted by* $i_1, i_2 < \cdots < i_k, A_{i_j} \subset X_{i_j}$ *is a point set of* X_{i_j}, *which makes the subset* $B = A_{i_1} \times A_{i_2} \times \cdots \times A_{i_k}$ *in* $X_{i_1} \times X_{i_2} \times \cdots \times X_{i_k}$ *be covered by the hypercube* $U(\epsilon), \epsilon > 0$ *in* $\mathbb{R}^k$, *then the subset* B *is called the* k*-dimension subspace* m*-object relation,* m *is the number of element in subset* B.

Note: In Definition 2.3, $U(\epsilon)$ refers to the k-dimensional hypercube of side length 2ϵ in $\mathbb{R}^k$, where $\epsilon > 0$ depends on the specific situation. For example, if we define $U(\epsilon) = (-\epsilon, \epsilon), \forall \epsilon > 0$ in the space of $\mathbb{R}^1$, then, there are one one-dimension subspace three-relation and two one-dimension subspace two-relations in Table 2 by simple calculation.

2.2 Properties of Subspace Multi-Relation

Theorem 2.1 *Subspace multi-relation is of downward compatibility. That is to say, if* B *is a* k*-dimension subspace* m*-relation, then for any* $k_1, m_1, 0 < k_1 < k, 0 < m_1 < m$, B *is also* k_1*-dimension subspace* m_1*-relation.*

Proof. Let $B = \{x_i | i = 1, 2, \cdots, m\}$ has the k-dimension m-relation. Given $\epsilon > 0$, then there exists a positive integer $k, \mathbb{R}^k \subset \mathbb{R}^n$, which makes the projection $x_i|_{\mathbb{R}^k} = (x_{i_1}, x_{i_2}, \cdots, x_{i_k})^T$ of x_i covered by the hypercube of $U(\epsilon), \epsilon > 0$. Thus for any subspace $\mathbb{R}_1^k$ with $\mathbb{R}_1^k \subset \mathbb{R}^k$, and any subset $\{x_i | i = i_1, \cdots, i_{m_1}\}$, $\| x_i - x_j \|_{\mathbb{R}^{k_1}} \leq \| x_i - x_j \|_{\mathbb{R}^k} < \epsilon$. So B is k_1-dimension subspace m_1-relation.

Theorem 2.2 *Subspace multi-relation is of upward aggregation. Assuming that all one-dimension subspace multi-relation in the set of objects S denoted by $s_i, i = 1, 2, \cdots, M$. If $s_{i_1 i_2 \cdots i_k} = s_{i_1} \cap s_{i_2} \cap \cdots \cap s_{i_k} \neq \emptyset, 1 \leq i_1 < i_2 < \cdots < i_k \leq M$, then $s_{i_1 i_2 \cdots i_k}$ is k-dimension subspace $m_{i_1 i_2 \cdots i_k}$-relation, where $m_{i_1 i_2 \cdots i_k}$ is the number of element in set $(s_{i_1 i_2 \cdots i_k}$.*

Proof. For any $x_i, x_j \in s_{i_1 i_2 \cdots i_k}, x_i = (x_{i_1}, \cdots, x_{i_k}), x_j = (x_{j_1}, \cdots, x_{j_k})$, To eliminate the influence of the different subspace calculation, we use the Manhattan distance which is normalized on dimension as following:

$$d_{\mathbb{R}^k}(x_i, x_j) = \sum_{s=1}^{k} |x_{i_s} - x_{j_s}|/k$$

As $x_i, x_j \in s_{i_1 i_2 \cdots i_k}$,then $|x_{i_s} - x_{j_s}| < \epsilon$,thus $d_{\mathbb{R}^k}(x_i, x_j) = \sum_{s=1}^{k} |x_{i_s} - x_{j_s}|/k < \epsilon$

Note: Here we assume that the $\epsilon > 0$ is uniform on different subspace in the proof, but the $\epsilon > 0$ can be different according to the actual problems.

3 Algorithm

We have given the definitions of subspace multi-relation and discussed its properties. In this section we will discuss the searching algorithm of subspace multi-relation and give some practical examples.

The subspace multi-relation is downward compatibility and upward aggregation, which allows us to adopt a bottom-up searching algorithm to obtain subspace multi-relation from low-dimension to high-dimension. Firstly, multi-relation is searched in one-dimension subspace. Secondly any two of one-dimension subspace multi-relation are used to compute intersection, and which being not empty is to form two-dimension subspace multi-relation. Thirdly, three-dimension subspace multi-relation could be found by computing the intersection of any two of two-dimension subspace relations. Finally, based on this process all the existing subspace multi-relations in attribute space will be found. This method is capable of avoiding the effect of the high-dimension disaster problem which is encountered during the direct search in the high-dimension space, and hence can greatly improve the computational efficiency.

Assume that there are m objects $x_i, i = 1, 2, \cdots, m$, which are expressed as $x_i = (x_{i_1}, \cdots, x_{i_n})^T \in \mathbb{R}^n, i = 1, 2, \cdots, m$ in the attribute space. Here we want to find the complex relations among the objects in attribute space based on subspace multi-relation model.

The main steps of subspace multi-relation searching algorithm are given as follow:

Step 1: Search one-dimension subspace multi-relation. Let $j = 1$, according to some clustering methods, we cluster the projection $\{x_{ij}, i = 1, 2, \cdots, m\}$ in one-dimension subspace $\mathbb{R}^j$ with each cluster covered by an interval of length 2ϵ. The result is denoted by $c^j_{(i)}, i = 1, 2, \cdots, m_j$, and let $s^j_{(i)} = \{x : x_{ij} \in c^j_{(i)}\}, i = 1, 2, \cdots, m_j$, which is belong to $\mathbb{R}^n$.

$j = j + 1$ until $j = n$.

Step 2: Search two-dimension subspace multi-relation. If

$$s^{j_1 j_2}_{(i_1 i_2)} = s^{j_1}_{i_1} \cap s^{j_2}_{i_2} \neq \emptyset$$

let

$$m^{j_1 j_2}_{(i_1 i_2)} = \mathrm{Num}(s^{j_1}_{i_1} \cap s^{j_2}_{i_2})$$

Then there exists a two dimension $m^{j_1 j_2}_{(i_1 i_2)}$-relation,$1 \leq j_1 < j_2 \leq n, 1 \leq i_1 \leq m_{j_1}, 1 \leq i_2 \leq m_{j_2}, i_1 < i_2$.

Step 3: Similarly, we can search for higher dimension subspace multi-relations. k=3; If

$$s^{j_1 j_2 \ldots j_k}_{i_1 i_2 \ldots i_k} = s^{j_1}_{i_1} \cap s^{j_2}_{i_2} \cap \ldots \cap s^{j_k}_{i_k} \neq \emptyset$$
$$1 < i_1 < i_2 < \cdots < i_k \leq m, 1 \leq i_t \leq m_i, t = 1, 2 \cdots, k$$

Then we say

$$s^{j_1 j_2 \cdots j_k}_{(i_1 i_2 \cdots i_k)} = s^{j_1}_{i_1} \cap s^{j_2}_{i_2} \cap \cdots \cap s^{j_k}_{i_k}$$

has k-dimension $m_{i_1 i_2 \cdots i_k}$-relation, where

$$m_{i_1 i_2 \cdots i_k} = \mathrm{Num}(s^{j_1 j_2 \cdots j_k}_{(i_1 i_2 \cdots i_k)})$$

k=k+1,If all of the intersection is empty, then stop searching.

Note 1: The clustering method of one-dimension space above can be suitably selected according to actual data such as k-means method or hierarchical clustering method;

Note 2: When dealing with big data, in order to control the number of subspace multi-relation, the number N of objects in which is asked to be greater than a given positive integer that is used to characterize how many objects are interesting, and parameter ϵ refers to the tightness of relation.

4 Examples

4.1 Experiment 1

In the department of mathematics of some college, the data set consists of 84 staff vertices, where each vertex has a total of 52 attributes including gender, education, job title, hobby, physical health, etc. The searching algorithm given above is used to find subspace multi-relation in data. The parameters ϵ is set to 0.5 and N (the number of staffs in subspace multi-relation) is 5.

The total 125 subspace multi-relations are found. By use of traditional graph model we can only obtain fully connective graph which lost a lot of information in data. That is because classical two-relation just indicate whether two objects have relation, but do pay attention to neither the simultaneously relation of how many objects nor in which subspace the relation happened. The more complex relation that exists in actual data can be expressed by the subspace multi-relation model, which can not only tell us how many objects have relation but also point out the attribute space which the relation exists in. In example given above, staffs link together differently because of schoolfellows, countrymen, or the same research field and so on.

4.2 Experiment 2

Enron was one of the most important companies in the U.S. energy industry. In 2001, the accounting fraud scandal of the Enron was exposed. Then the Federal Energy Planning Board took e-mail communications between Enron employees posted online [17] for publicity and academic research. Here, we want to discover some interesting groups based on these email. From the body content of those emails which happened before the fraud scandal we extract 2062 different people as vertices and set 144 staff's mailbox as the attributes of vertices. Let ϵ be 0.5 and N be 60, we finally find 275 subspace multi-relations by our searching algorithm.

It is very interesting that the highest dimension of subspace in all subspace multi-relation is 8 with 86 mailboxes, referring to 8 people including Chief Operating Officer, Director of Risk Management, Vice President and Chief, Online President of Enron, Enron CEO in North America, and other two important figures, who all are suspected of fraud, which tell us Maybe the event come from people's talking.

5 Subspace Multi-Relation And Vector Hyper-Graph

As two-dimension relation corresponds to the graph model, subspace multi-relation corresponds to the hyper-graph model. Recently studies on hyper-graph model have mainly focused on how to expand the existed properties in graph theory into hyper-graph, such as hyper-path, hyper-chain, graph partitioning and so on [9-13]. Furthermore, hyper-graph models do not indicate that the multi-relation exists in different subspaces.

Here we propose the concept of vector hyper-graph corresponding to subspace multi-relation.

Definition 5.1 *let $X = \{x_1, x_2, \cdots, x_s\}, x_i \in \mathbb{R}^n$ be a finite set. A vector hyper-graph of X is denoted by $G = (E_1, \mathbb{R}^{k_1}), \cdots, (E_p, \mathbb{R}^{k_p})$ where E_i is the finite collection of subset of X, such that:*

$$(1) E_i \neq \theta, i = 1, 2 \cdot\cdot\cdot, p$$

$$(2) \bigcup_{i=1}^{p} E_i = X, \mathbb{R}^{k_i} < \mathbb{R}^n, 1 \leq k_i \leq n, i = 1, 2 \cdots p$$

In the vector hyper-graph G, x_i is called a vertex, $(E_i, \mathbb{R}_i^k)$ is called vector hyper-edge. There exists a natural connection between the vector hyper-graph and subspace multi-relation. Assume that a vector hyper-graph X has a vector hyper-edge $(E_i, \mathbb{R}^{k_i}) = (x_{i_1}, x_{i_2}, \cdots, x_{i_{m_i}}, R^{k_i})$, then the objects $x_{i_1}, x_{i_2}, \cdots, x_{i_{m_i}}$ has multi-relation on subspace $\mathbb{R}^{k_i}$, which means k_i-dimension m_i-relation, and vice versa. Hence we can use the vector hyper-graph to express subspace multi-relation and study the complex relation in system.

6 Conclusions

We discuss the method to express the complexity of a system by a subspace multi-relation model. Based on the attribute space of objects, we establish the rigorous definition of subspace multi-relation. The traditional graph model or complex network is a special case which is equivalent to the one-dimension two-relation. Moreover, an algorithm is given to search for the subspace multi-relation and its effectiveness is verified by real data. Finally, vector hyper-graphs related to multi-relation models are discussed. The equivalence of the subspace multi-relation and the vector hyper-graph is pointed out, which provides a new method to study system complexity through vector hyper-graphs.

References

[1] Nicholas W. Watkins and Mervyn P. Freeman. (2008), "Natural Complexity", *Science*, Vol.320, No.5874, pp.323-324.

[2] Yi Lin. (2008), *Systemic Yoyos:Some Impacts of the Second Dimension*, CRC Press.

[3] Pavel Kabat. (2012), "Systems Science for Policy Evaluation". *Science*, Vol.336, No.6087, pp.1398.

[4] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. (2012), "Detecting Causality in Complex Ecosystems", *Science*, Vol.338, No.6106, pp.496-500.

[5] J. Leskovec, J. Kleinberg and C And Faloutsos. (2005), "Graphs over time: densification laws, shrinking diameters and possible explanations", *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

[6] M.E.J. Newman. (2003), "The structure and function of complex networks", *SIAM Rev*, Vol.45, No.2, pp.167-256.

[7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Huang. (2006), "Complex networks: structure and dynamics", *Phys. Rep*, Vol.424, No.4, pp.175-308.

[8] L. da, F. Costa, F.A. Rodrigues, G. Travieso And P.R.U. Boas. (2006), "Characterization of complex networks: a survey of measurements", *Adv. Phys*, Vol.56, No.1, pp.167-242.

[9] Claude Berge. (1989), *Hypergraphs: Combinatorics of finite sets*, North-Holland.

[10] D.S.Johnson and H.O.Pollak. (2006), "Hypergraph planarity and the complexity of drawing venn diagrams", *Journal of graph theory*, Vol.11, No.3, pp.309-325.

[11] Vitaly I.Voloshin. (2009), *Introduction to Graph and Hypergraph Theory*, Nova Science Publishers, Inc.

[12] Xie Zheng, Yi Dong-Yun, Ouyang Zhen-Zheng and LI Dong. (2012), "Hyperedge communities and modularity reveal structure for documents", *China. Phys. Lett.*, Vol.29, No.3.

[13] Dominique LaSalle and George Karypis. (2013), "Multi-threaded Graph Partitioning", *27th IEEE International Parallel & Distributed Processing Symposium.*

[14] Charu C.Aggarwal, Joel L.Wolf, Philip S.Yu, Cecilia M. Procopiuc and Jong Soo Park. (1999), "Fast algorithms for projected clustering", *ACM SIGMOD Record*, New York, Vol.28, No.2, pp.61-72.

[15] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos and Prabhakar Raghavan. (2005), "Automatic subspace clustering of high dimensional data", *Data Mining and Knowledge Discovery*, Vol.11, No.1, pp.5-33.

[16] Kriegel Hans-Peter, Kroger Peer and Zimek Arthur. (2009), "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering", *ACM Transactions on Knowledge Discovery from Data*, Nwe York, Vol.3, No.1, pp.1-58.

[17] Enron email dataset, http://www.cs.cmu.edu/ enron/

Corresponding author
Dongyun Yi: dongyunyi@sina.com