Journal of Turkish

Operations Management

The inference of complicated networks by mutual information

Hajar Franoudkia[1*],

[1]Department of Business Administrations, Başkent University, Ankara
e-mail: hajerfarnoudkia@baskent.edu.tr, ORCID No: https://orcid.org/0000-0001-9201-663X
*Corresponding Author

Article Info

Article History:
Received: 28.02.2023
Revised: 11.05.2023
Accepted: 26.05.2023

Keywords
Conditional dependence,
Gaussian copula,
Sahnnon entropy,
Mutual information

Abstract

Unsupervised machine learning affords a general idea about complicated data using a graphical representation of networks by nodes and edges to provide a better and easier understanding of the data. The existence of an edge between two entire nodes is determined by their relationship in terms of any kind of dependence i.e., conditional dependence, linear and non-linear, directed or undirected. This study tries to show the accuracy of a non-parametric approach i.e., mutual information (MI) on a real data set named by the *Rochdale* data that is composed of eight factors effected on women's economic activity by comparing with some methods such as reversible jump Markov Chain Monte Carlo (MCMC) and birth-death MCMC those tried to detect the conditional dependence between the variables. As a result, MI is not only a very simple but also a very accurate method in the inference of data with complexities.

1. Introduction

Unsupervised machine learning is a powerful tool that discovers patterns and relationships in data without the need for labeled training data. A common approach in unsupervised learning is to use mutual information, a measure of statistical dependence between two variables, to detect patterns and structure in data. By analyzing the mutual information between different features or variables in a dataset, the hidden relationships and dependencies can be detected that may not be immediately apparent. On the other side, the structure of complex systems is frequently shown and analyzed using graphical representations of networks. While edges in a network reflect the connections or links between nodes, nodes in a network represent entities or objects. Networks can be graphically represented in a variety of ways, such as with node-link diagrams, adjacency matrices, and force-directed layouts.
As an explanation of the node-link diagrams, suppose a $(n \times p)$-dimensional data set where p is the number of columns (variable) and n is the number of rows (sample). This data will be represented graphically as a network with p nodes to embody the variables. Let's call E the set of edges that can be written as a set of $\{(i,j); i,j = 1, \dots, p\}$ which Y_i and Y_j are connected by a un/directed edge. In the directed graphs, the relationship matric is not symmetric means that the existence of an edge between Y_i to Y_j does not imply an edge from Y_j to Y_i. The study of directed networks, also known as directed graphs, has a long history in mathematics, computer science, and other fields. There are more than six thousand papers that scrutinized the directed edges between the variables since 1953 started by Harary and Norman (1953). There are still many unanswered topics in the study of directed networks today, and new applications are constantly being developed. On the other side, undirected graphs are a type of graph in which the edges between nodes are not directed, meaning that they do not have a specific direction associated with them means if there is an edge from Y_i to Y_j , there is also an edge from Y_j to Y_i . Undirected graphs have been studied extensively in graph theory and have many applications in various fields, such as computer science, social network analysis, and transportation planning. They are often used to model relationships between objects or entities where the direction of the relationship is not important, such as in a social network where the connections are bidirectional. There are around three thousand recorded articles based on undirected graphs started by Borowiecki, (1947). Among these studies, some researchers investigated conditional dependence when each variable is written as a regression equation based on another variable given the remaining variables. (Dobra and Lenkoski, 2011, Mohammadi and Wit, 2015, Farnoudkia and Purutcuoglu, 2019). In continuation, the undirected and directed graphs are implemented in the presentation of time-related data, as well. (Abegaz and Wit,

2013). The adjacency matrix, on the other hand, is a binary matrix with the dimension of $p \times p$ where p is the number of variables (nodes) in which 1 stands for the related and 0 for non-related corresponding variables.
Apart from all discussed until here, there is a very general and non-parametric tool that can measure the strength of the connection between two variables called mutual information (MI) shown by $I(X_i, X_j)$ introduced by McGill in 1954. This measure is a non-negative entity with zero value for the independence case. MI is a symmetric measure like a correlation coefficient and is able to catch the dependence even in a non-linear case, unlike the correlation coefficient. The definition and details of MI will be stated in Section 2. Furthermore, two alternatives are explained briefly in Section 3. Finally, the accuracy of MI will be stated by two accuracy measures and compared with two alternatives in the Application section for a real data set to detect the symmetric (undirected) reletionship between every two variables.

2. Mutual information

Mutual information is a function of the univariate and bivariate Shannon entropy (Sahnnon ,1949)shows the common information of two processes or variables as below:

$$I(X_i, X_j) = H(X_i) + H(X_j) - H(X_i, X_j) = H(X_i) - H(X_i|X_j) = H(X_j) - H(X_j|X_i) \quad (1)$$

Where H stands for Shannon entropy.
For discrete cases, if variable X_i takes the value of $\{x_{i1}, \dots, x_{ik}\}$ with the probability of $p_1, \dots, p_k$ that is a multinomial distribution which is the mother of the discrete distributions, then the Shannon entropy is defined by $H(X_i) = -\sum_{r=1}^{k} p_r \log(p_r)$ shows the uncertainty corresponding to X_i.
For the bivariate case, suppose there is another variable X_j taking the values of $\{x_{j1}, \dots, x_{jl}\}$ with the joint probability function with X_i as $P_{X_i,X_j}(x_{ir}, x_{js})$ for $r = \{1, \dots, k\}$ and $s = \{1, \dots, l\}$. The definition of the joint Shannon entropy of X_i and X_j is

$$H(X_i, X_j) = -\sum_{r=1}^{k} \sum_{s=1}^{l} p_{X_i,X_j}(x_{ir}, x_{js}) \log(p_{X_i,X_j}(x_{ir}, x_{js})) \quad (2)$$

For the dependent case, $H(X_i, X_j) = H(X_i) + H(X_j)$ and then $I(X_i, X_j) = 0$.
MI is a symmetric measure that assures us to use it in the inference of undirected networks.
To use it in directed networks some other measures are introduced like transfer entropy (TE) (Schreiber, 2000) that is for time data and also K-dependence coefficient defined as $K(X_i : X_j) = \frac{I(X_i,X_j)}{H(X_i)} \in [0,1]$ (Kong, 2007).

3. Some Alternatives

As mentioned before, this study aims to provide a very simple way to detect the dependent variables to create an undirected graph of the data. We employ the MI on a real data set composed of binary values due to its better definition for discrete random variables. To compare MI accuracy, some alternatives such as Reversible Jump Markov Chain Monte Carlo (MCMC) and Birth-Death MCMC try to estimate the inverse covariance matrix to determine the conditional dependence between the variable after transforming the data into Gaussian by Copula. The details can be found in the study of Farnoudkia and Purutçuoğlu (2017). In this section, the two alternatives will be explained in detail that try to construct an undirected graph for the data set.

3.1 Reversible Jump Markov Chain Monte Carlo

The Bayesian approach is a very well-known method provides a more accurate estimation of the parameters, where the estimation most of the time can be written in terms of the prior estimator and maximum likelihood estimator. On another side, MCMC methods can offer an estimation of the parameters based on an iterative algorithm in which, the estimation of each iteration is not worse than the previous one. The precision matrix is responsible for the conditional dependence between two normally distributed variables, where the number of non-zero elements can change in each iteration when the precision matrix is estimated by MCMC method. That is why RJMCMC is proposed as a suitable algorithm that is compatible with the dimension-varying problem. The values of the estimated precision matrix from this method are responsible for the conditional dependence between corresponding normally-distributed variables in which the zero elements indicated the conditional independence.
If the data is not normally distributed, the copula can transform the data into a normally distributed one using the inverse of the empirical cumulative distribution of the variables in the Gaussian Copula function. More details are referred to in Kojadinovic and Yan (2010).

3.2 Birth-Death Markov Chain Monte Carlo

BDMCMC is another method based on the continuous-time approach which estimates the precision matrix when the data is normally distributed and the parameter dimension is varying like RJMCMC. In this algorithm, both birth rate and death rate are calculated by the Poisson process. So in each iteration, the zero elements of the precision matrix are born with a birth rate and the non-zero elements would die by the death rate. Then, the precision matrix is updated at the end of each iteration. The choice of the birth and death rates determines the birth-

death process and is made in such a way that the stationary distribution is precisely the posterior distribution of interest. Contrary to the RJMCMC approach, the moves between models are always accepted, which makes the BDMCMC approach extremely efficient and fast. More details can be found in Mohammadi and Wit (2017).

4 Application

The only data that is used in this study is a real set named the *Rochdale* data. To calculate the accuracy of the proposed method and some alternatives the adjacency matrix is estimated and evaluated by two measures as follows.

- F_1-score which is equal to 1 for the best case and 0 for the worst with the following formula.

$$F_1\text{-score}= \frac{2TP}{2TP+FP+FN} \tag{3}$$

where TP stands for true positive edges, FP for the false positive edges, FN for the false negative, and TN for the true negative edges. These are the four elements of the confusion matrix where the estimation is binary that is suitable fort his study.

- Mathew correlation coefficient (MCC) lying between -1 and +1, one for the most accurate case and -1 when there is a complete disagreement between the prediction and observed values and zero for the random case as follows.

$$\text{MCC}= \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+Fp)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

4.1 The Rochdale data

The Rochdale data is a set of eight binary economic factors which is firstly modeled by the log-linear model in the study of Whittaker (2009). The data is gathered by 665 persons answering eight YES/NO questions. The questions (variables) are labeled by letters as *a*: 1 for economically active wife, *b*: 1 if the wife is older than 38, *c*: 1 if the husband is unemployed, *d*: 1 if the family has more than 4 children, *e*: 1 of the education level of the wife is more than high school, *f*: 1 if the education level of husband is more than high-school, *g*: 1 for the Asian origin, and *h*: 1 of other household member is working. The data is represented conveniently in Table 1. where each cell represents one of the 2^8 the possible combination of 0 and 1. For instance, the first cell shows that 5 persons out of 665 answered all eight questions as zero. For instance, the 9th row and the 13th column which is 57 mean that 57 persons out of 665 answered the questions by, $a = 0, b = 0, c = 0, d = 0, e = 1, f = 1, g = 0, h = 0$.

Table 1. The Rochdale data

5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0
8	0	11	0	13	0	0	0	3	0	1	0	26	0	1	0
5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0
17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0
1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3	2	0	23	4	0	0	22	2	0	0	57	3	1	1
5	1	0	0	11	0	1	0	11	0	0	0	29	2	0	0
3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0
0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0
2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The true network is composed of 14 edges as *ac, ad, ae, ag, bd, be, bh, ce, cf, cg, dg, dh, ef, fg*. To verify the accuracy of the proposed method, the adjacency matrix (an $8 \times 8 -$ dimensional binary matrix) is estimated by MI for this study and the other two methods from the study of Farnoudkia and Purutcuoglu (2017). Table 2 shows the accuracy of different methods in two measures.

Table 2. F_1-score and MCC measures of three differenet methods

	True Network	**MI**	**RJMCMC**	**BDMCMC**
TP	14	14	12	11
TN	14	14	14	10
FP	0	0	0	4

FN	0	0	2	3
F_1-Score	1	1	0.92	0.76
MCC	1	1	0.87	0.50

5. Discussion and Conclusion

This study aims to suggest the use of the MI of two variables as criteria for any relationship between them which is one of the main aspects of unsupervised machine learning methods. MI is a non-parametric measure and easy to apply but the point is that this measure is designed for discrete random variables. Mutual information can be used in stock market analysis to identify the relationships between different variables and their impact on stock prices. By calculating the mutual information between various economic indicators and stock prices, we can determine which variables have the strongest influence on the market (Farnoudkia and Purutçuoğlu, 2020). However, for continuous random variables, the method of binding should be implemented first. There are some other methods like Gaussian Copula transformation as well. The application section proves the high accuracy of MI for the binary data set by comparing it with the true network and also by two alternative methods' accuracy. In this study, the graph is an undirected graph and the true graph is determined by conditional dependence which does not necessarily coincide with dependence. In the future study, other types of data will be used as well as other types of relationships. In a nutshell, MI is superior to be used at least for the independent variables because a zero MI implies linear and non-linear independence.

Conflict of Interest

The author declares that she has no conflicts of interest regarding the publication of this article. She received no financial support or funding for this research. The author wishes to express her sincere gratitude to the editor-in-chief and other members of the journal for their valuable feedback and assistance throughout the review process. Their contributions have greatly improved the quality and rigor of this work.

References

Abegaz, F., & Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, *14*(3), 586-599. https://doi.org/10.1093/biostatistics/kxt001.

Borowiecki, M. (1947). On the problems of isomorphism and construction of oriented graphs. In *Colloquium Mathematicum* (Vol. 1, p. 1). Editions Scientifiques de Pologne. https://doi.org/10.4064/cm-1-1-37-50.

Dobra, A., & Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. The Annals of Applied Statistics, 5(3), 969-993. https://doi.org/10.1214/10-AOAS439

Farnoudkia, H., Purutçuoğlu, V. (2020). Application of r-vine copula method in Istanbul stock market data: A case study for the construction sector. *Journal of Turkish Operations Management*, 4:509-518. https://dergipark.org.tr/tr/pub/jtom/issue/59336/851947.

Harary, F., & Norman, R. Z. (1953). *Graph theory as a mathematical model in social science* (No. 2). Ann Arbor: the University of Michigan, Institute for Social Research. https://doi.org/10.1017/s1373971900075089

Kojadinovic, I., & Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, *34*, 1-20. https://doi.org/10.18637/jss.v034.i09.

Kong, N. (2007). An entropy-based measure of dependence between two groups of random variables. *ETS Research Report Series*, 1, i-18. https://files.eric.ed.gov/fulltext/EJ1111559.pdf.

McGill, W. (1954). Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 93-111. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1057469.

Mohammadi, A., & Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. https://doi.org/10.1007/s11222-014-9523-7

Mohammadi, R., & Wit, E. C. (2017). An Introduction to the BDgraph for Bayesian Graphical Models. https://pure.uva.nl/ws/files/25409351/1712_Crop_.pdf.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, *85*(2), 461. https://doi.org/10.1103/PhysRevLett.85.461.

Shannon, C. E., & Weaver, W. (1949). A mathematical model of communication. *Urbana, IL: University of Illinois Press*, *11*, 11-20. https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing. https://doi.org/10.1002/9780470744639.