From Association Analysis to Causal Discovery

[Extended Abstract]

Jiuyong Li
School of Information Technology and Mathematical Sciences
University of South Australia, Australia

Association analysis is an important technique in data mining, and it has been widely used in many application areas [6]. However, associations found in data can be spurious and do not reflect the 'true' relationships between the variables under consideration. For example, it is easily for hundreds or thousands of association rules to be generated even in a small data set, but most of them could be spurious and have no practical meaning [11, 21, 22]. This has hindered the applications of association analysis to solving real world problems. While the development of efficient techniques for finding association patterns in data, especially in large data sets, is well underway, the problem for identifying non-spurious associations has become prominent.

Causal relationships imply the real data generating mechanisms and how the outcome would change when the cause is changed, so finding them has been the ultimate goals of many scientific explorations and social studies [18]. The gold standard for causal discover is randomised controlled trials (RCTs) [4, 16]. However, a RCT is infeasible in many real world applications, particularly in the case of high dimensional problem of a large number of potential causes. As part of the efforts on causal discovery, statisticians have studied various methods for testing a hypothetical causal relationship based on observational data [16]. However, these methods are designed for validating a known candidate causal relationship and they are incapable of dealing with a large number of potential causes either.

Although an association between two variables does not always imply causation, it is well known that associations are indicators for causal relationships [7]. Therefore a practical approach to causal discovery in large data sets could start with association analysis of the data.

A question is then whether we can filter out associations that do not have causal indications. Note that this objective is different from that of mining interesting associations [9, 20] or discovering statistically sound associations [5, 21] because interestingness criteria do not measure causality and a test of statistical significance only determines if an association is due to random chance. We have integrated two statistical methods for testing a hypothetical causal relationship respectively with efficient association analysis techniques for automatic discovery of causal relationships in large data sets.

One solution is to test partial association [1, 14] instead of marginal association in data. When we study the relationship between two variables, the effects of other variables should be considered since they may have influence on the discovery of the relationship between the two variables under study. A partial association test considers the association of two variables when the other variables are 'controlled', i.e. while the values of other variables are known to be the same. If the association between the two variables is non-persistent in most control conditions, it is not a causal indicator. However, the complexity of a partial association test is exponential to the number of variables. We have designed an efficient algorithm for such a test and also integrate the test into an association rule mining process [8]. Experiments with synthetic and real world data have shown that the method is efficient for detecting causal relationships in large data sets.

Another solution is to conduct a retrospective cohort study in data [3] to validate a causal relationship. A cohort study is a type of observational studies used in medical and social research, where RCTs are practically impossible, to infer risk factors. It normally follows two groups of individuals, who share common characteristics but differ in regard to a certain factor of interest to infer how the factor causes an outcome. When the information about individuals has been collected sufficiently in data, a retrospective cohort study can be used to infer risk factors. In a social or medical study, a causal hypothesis needs to be presented prior to the data being selected for a cohort study. Therefore cohort studies have not been applied to causal discovery in large scale data when candidate causal relationships are unknown. In our research [10], we combine association analysis with cohort studies to generate causal hypotheses from data sets with a large number of potential causes and to test the hypotheses in one causal discovery process. This method has shown promising results in some real world data sets and its efficiency has also been demonstrated by experiments.

One common characteristic of the two proposed methods is that they are capable of finding causal factors of combined variables, which are impossible to be uncovered by other existing methods, e.g. those using the causal Bayesian network scheme [2, 12, 13, 17, 15, 19]. The two proposed methods provide efficient alternatives to the causal Bayesian network approaches for the discovery of causal relationships in large data sets.

MLSDA 2013, Dunedin, New Zealand

1. REFERENCES

[1] M. W. Birch. The detection of partial association, I: the 2x2 case. *Journal of the Royal Statistical Society*, 26(2):313–324, 1964.

[2] G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

[3] A. M. Euser, C. Zoccali, K. J. Jager, and F. W. Dekker. Cohort studies: prospective versus retrospective. *Nephron Clinical practice*, 113(3), 2009.

[4] I. Guyon, D. Janzing, and B. Schölkopf. Causality: Objectives and assessment. *Journal of Machine Learning Research - Proceedings Track*, 6:1–42, 2010.

[5] W. Hämäläinen. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge Information Systems*, 32(2):383–414, 2012.

[6] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, third edition, 2011.

[7] A. B. Hill. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300, 1965.

[8] Z. Jin, J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang. Discovery of causal rules using partial association. In *Proceedings of IEEE International Conference on Data Mining*, pages 309–318. IEEE Computer Society, 2012.

[9] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610–626, 2008.

[10] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, and B. Sun. Mining causal association rules. In *Proceedings of ICDM Workshop on Causal Discovery (CD)*, 2013.

[11] G. Liu, H. Zhang, and L. Wong. Controlling false positives in association rule mining. *Proceedings of VLDB Endowment*, 5(2):145–156, 2011.

[12] M. H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248, 2010.

[13] M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6A):3133–3164, 2009.

[14] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

[15] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[16] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2 edition, 2001.

[17] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2-3):163–192, 2000.

[18] P. Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11:1643–1662, 2010.

[19] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, USA, second edition, 2000.

[20] P. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293 – 313, 2004.

[21] G. I. Webb. Discovering significant patterns. *Machine Learning*, 71(1):1–31, 2008.

[22] G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71:307–323, 2008.