2010 International Conference on Business and Economics Research
vol.1 (2011) © (2011) IACSIT Press, Kuala Lumpur, Malaysia

Causal Knowledge-Driven Approach For Stock Analysis

Alireza Khorram
Faculty of information science and technology
Multimedia University
Melaka, Malaysia
Khorram83@gmail.com

Dr.Cheah Wooi Ping
Faculty of information science and technology
Multimedia University
Melaka, Malaysia
wpcheah@mmu.edu.my

Liew Tze Hui
Faculty of information science and technology
Multimedia University
Melaka, Malaysia
thliew@mmu.edu.my

Abstract—**Because of the causal nature of stock price changes, Causal network can be used to model the relationships between stocks. Unfortunately, there are very few works on the application of causal knowledge-driven approach in stock analysis; in particular, learning causal models from data. In this study, we introduce learning Bayesian networks from data as an applicable model for representing and reasoning about stock market changes. As a case study, a portion of 100 Malaysian stock securities (FBM100) data is used. The Bayesian model has been used to perform both diagnosis and prognosis on the learnt model. A methodology is proposed to apply causal knowledge-driven approach for stock analysis using Bayesian networks. We have adopted Tetrad IV and Genie 2.0 as a Bayesian network learner and inference engine. As of our objective, the probable price of some selected stocks has been predicted. Prediction has been performed as sign prediction (increase or decrease) and probable value prediction. The results show reasonable degree of satisfaction since they are tested against real world scenarios. However, there are some limitations which have been discussed.**

Keywords: Knowledge-driven approach, Causal reasoning, Bayesian network, Directed acyclic graph and conditional independence.

I. Introduction

Today, large amount of data exists in various areas of science and economic enterprises, which can be used as a rich resource for knowledge discovery. To realize, interpret and use these data, data mining approach has been proposed. Data mining and knowledge discovery technology is suitable for overcoming the limitations of the traditional methods and to find out useful information from data. Knowledge Discovery and Data Mining (KDD) refers to an area which comprise of many fields of studies and concentrate on developing and improving methodologies which can be used for extracting valuable knowledge from data. [1]

Due to the causal nature of stock data, we are facing with causal knowledge which is a kind of knowledge considering cause and effect impacts between domain variables. The method or technique that enables us to represent or reason about this kind of knowledge is called causal knowledge-driven approach. In general, Knowledge representation refers to formal reconstruction of knowledge and its implementation [2]. The reasoning about knowledge refers to computational methods which can extract new knowledge from existing one [2]. In economic jargons, stock analysis is the process of determining and tracing patterns which have existed in stock price changes. It determines the probable value of the stock price in the next slice of time. For predicting the stock market price changes, various methods have been utilized in literature.

For modeling, representing and reasoning about causal knowledge, two major frameworks are existed. These frameworks are fuzzy cognitive map (FCM) and Bayesian network (BN).

Bayesian network which can be viewed as intersection of statistics and artificial intelligence is a well established method for probabilistic causal reasoning. It uses graphical structure to represent causal relationships and update belief given new information. Bayesian network is helpful when no experiment about the effects of change(s) of one data variable is available [3]. Moreover, it is useful in situations which plagued with uncertainty.

A. *Problem Definition*

Generally, the nature of relations between stock variables are causal and a causal model is used to capture these relationships. The first step for constructing a causal model is to identify the structure of the model which can be represented as Bayesian network.

In stock analysis, most cases described in the literature use knowledge engineering approach, in which knowledge experts identify the relationships between data elements. This approach is prone to some problems when we are facing with large number of variables and complexity is high. Therefore, a methodology for learning the structure of the network from stock data is needed. The proposed method should be able to learn causal model from data without intervention of knowledge experts.

B. *Objectives and scope of the study*

Our main objective of this study is to use causal knowledge-driven approach for stock analysis. A causal knowledge model in Bayesian network is learnt from stock data. The Bayesian model is then used to perform both diagnosis and prognosis reasoning on the learnt model.

The second objective of this study is to predict the probable price of the stock for next day.

We will focus on using Bayesian network as a causal model for stock data. Our main concern is to use Bayesian learning approach to extract a causal model from stock data and identify hidden relationships between these variables which are unseen from knowledge experts and can affect the stock price.

II. BAYESIAN NETWORKS

A. *basics of bayesian networks:*

Bayesian networks are directed acyclic graphs (DAG) which each node represents random variables and each connection between two nodes represent causal dependency between the two variables [4]. Bayesian networks have come to existence as a framework for representing and reasoning about uncertain knowledge [5]. An example of BN for lung cancer problem is as follows [6].

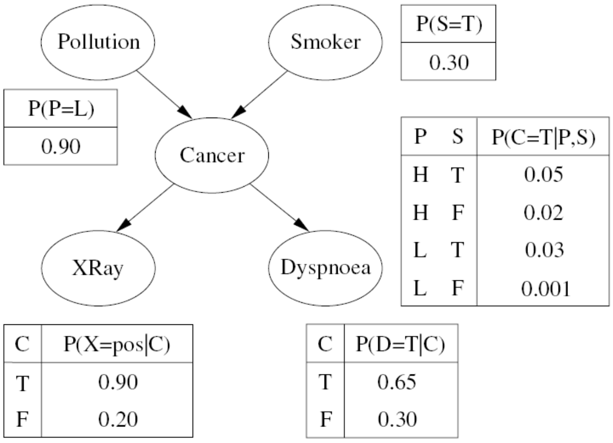

Figure 1. A BN for lung cancer problem

Bayesian networks applied in wide variety of problems. It is extensively and successfully exploited in bioinformatics [3]. Another successful area which Bayesian networks are used is medical diagnosis [7]. In stock analysis, Bayesian network is used for causal reasoning. But in most cases, it is constructed using knowledge engineering approach.

In literature, there exist numerous papers which applied Bayesian networks for causal reasoning and knowledge representation. It is possible to group these papers to three main categories.

- The first category is the one that use knowledge experts' opinions in constructing Bayesian network. This approach is called knowledge engineering approach.
- The second category uses Bayesian methods for classification purposes.
- The third category includes the papers that use learning Bayesian network from data.

Reference [8] tried both knowledge engineering and learning methods in medical area (cardiovascular Bayesian network).

B. *bayesian network software tools:*

Over past 20 years, many researches have been done and still being done on Bayesian networks and many algorithms have been developed. This is accompanied by rapid growth of Bayesian software tools.

Reference [9] has collected a list of software packages which has been developed for graphical models and Bayesian networks during a number of years.

Reference [6] reviewed Murphy's software list and described some of the major software packages which have the most functionality or have particular feature. The most noticeable Bayesian tools which are commonly mentioned as a good BN packages in the literature are: Hugin, GeNie, Tetrad IV, Netica, CaMML and BNT. Most of these packages have advanced graphical user interface (GUI) features which make them convenient and easy to use (except CaMML which is a powerful package but works on a text based mode).

Tetrad is Bayesian software which is developed by Peter Spirtes, Clark Glymour and Richard Scheines in Carnegie Mellon University. The authors are the ones who created PC algorithm. Version 2 of Tetrad was the first commercially causal discovery program [6]. The latest version of Tetrad is IV (version 4) which is freely available and supports main causal discovery algorithms. Moreover, it is written in java runtime environment, and has very powerful GUI and also, it is easy to use. It provides many options to the user in the case of choosing implemented techniques or algorithms. However, there exist some issues in using Tetrad. It does not permit users to export the graphs and CP tables to other standard format. Moreover, there is not any print or export facility in the software.

GeNie is a development environment which is developed by university of Pittsburgh and supports decision networks and Bayesian networks. GeNie performs inference reasonably good and supports diverse algorithms [6].

C. *Bayesian networks learning algorithms*

There are two different ways which researchers encountered with learning problem. One of these methods is dependency analysis while the other is search-and-scoring approaches [10].

In dependency analysis approach, Bayesian network is viewed as a network which depicts conditional independence relations among random variables. So, the approach attempts to construct BN from conditional independence relations which are obtained from given data. Reference [11] reviewed some noticeable algorithms which are designed for learning Bayesian network from data.

In search-and-scoring approach, Bayesian network is viewed as a structure which encodes joint probability distributions of variables. A measure is used (Bayesian, MDL or KL entropy scoring function) as a criteria for finding out the best BN which maximizes the used measure and best fits the data. [11].

The two general approaches for learning Bayesian networks are compared in [11] and it is shown that scoring – and –search methods in learning BNs have certain advantages in compared with dependency analysis methods [11].

It is also mentioned that for large number of variables, the constraint based methods are more efficient. On the contrary, since the score-based algorithms search whole model space to find the optimal model, they are more efficient and more accurate when the sample size is small and data is noisy [11].

According to [11], we have decided to use constraint-based methods for our research and between existed algorithms which belongs to this category; PC algorithm [12] seems to be more suitable for our research. Moreover, PC is implemented in Tetrad and is available for free. The algorithm starts with a complete undirected graph whose nodes are domain variables. Then the algorithm try to thin the graph based on d-separation concept with zero-order, one-order (and so on) conditional independence relations. After the graph is reduced in number of edges and no more relationships can be removed, the algorithm orients the remaining graph with a pre-defined pattern. At the end, the remaining edges which are still undirected will be oriented so that no directed cycle occurs in the graph.

D. Bayesian network inference mechanisms

Any probabilistic inference system aims to calculate the posterior probability distribution for a set of query variables when some observed event(s) are given [13].

There are two kinds of inference mechanisms in Bayesian networks. These are exact inference and approximation inference. In the following, some important algorithms of these categories are mentioned [14].

A very popular algorithm which is a kind of exact inference algorithms is Clustering algorithm. Clustering algorithm is the fastest known exact algorithm for belief updating in Bayesian networks. It was originally proposed by [18].

The clustering algorithm works in two phases: (1) compilation of a directed graph into a junction tree, and (2) probability updating in the junction tree.

The clustering algorithm is GeNIe's default algorithm. Only when networks become very large and complex, the clustering algorithm may not be fast enough.

Clustering algorithms (also known as join tree algorithms) need O(n) for time. For this reason, these algorithms are widely used in commercial Bayesian network tools.

The other type of inference algorithm is approximate inference in Bayesian network. The noticeable example of this group is MCMC algorithm.

MCMC (Markov chain Monte Carlo) generates each event by making a random change to the preceding event. The next state is generated by randomly sampling a value for one of the non evidence variables Xi, conditioned on the current values of the variables in the Markov blanket of Xi [16]. MCMC algorithm is implemented in CaMML software. In this study, we have used clustering algorithm for inference. As we mentioned, it is default inference algorithm in GeNie 2.0. There are many other algorithms which are implemented in GeNie 2.0 such as AIS sampling, polytree, EPIS sampling, logic sampling, backward sampling and likelihood sampling. It should be noted that Tetrad IV does not provide any inference mechanism and it is intended to be used in structure learning.

III. METHODOLOGY

In this study, we want to apply causal knowledge-driven approach for stock analysis using data supplied by a knowledge expert. 100 main Malaysian stocks (FBM 100) are considered as a case. We will use Bayesian network to represent the causal model to capitalizing on its strength in diagnosis and prognosis. Because of powerful capabilities and popularity, Tetrad IV and GiNIe 2.0 are used as a Bayesian network learner and inference engine since they are able to do structure and parameter learning.

The methodology is composed of four steps as follows:

1. Pre-process the data into proper format, suitable for causal knowledge recovery.

We have categorized each stock based on its domain. These categories (which are called sectors in economy) are based on the formal sectors which are existed in Malaysian stock market. They are composed of Constructions, Finance, Industrial products, Consumer products, Technology, Plantations, Properties and Trading and Services. Moreover, we have disceretized the data into three-levels (for sign prediction) and 9-levels format (for value prediction).

2. Discover the relationships between variables (data elements) using Bayesian network and construct the model.

This step composed of two parts. First part is structural learning which identifies the structure of the causal network, as mentioned before; PC algorithm is used for structure learning. The second part is parameter learning and is performed after the structure of the network is learned. In this step, the conditional probability table (CPT) for each node is identified. This table is our criterion for judging about how much is the probability of the influence of one node to another. Generally, the CPT for a variable represents the conditional probability of the variable given every combination of the values of its parents.

3. After the causal model has been constructed, we will perform both prognosis and diagnosis on the stock model.

4. Finally, test the accuracy of the model against the real world scenario.

For testing purposes, one fifth of the given data were used. We have fixed a selected node and traced the effect of the changes of the other nodes on the selected nodes, based on the data on hand.

IV. EXPERIMENTAL RESULTS

For the first experience, we have applied our proposed methodology within each sector data. Our aim is to determine whether the algorithm is capable of indentifying the basic relationships which exist between stocks within the sector and the sector indicator. The result shows that the reltionships are successfully identifed. (figure 2)

On our second experiment, we have applied our methodology between sector indicators. Because the sector indicators are the stocks which exist from the beginning days of Bursa Malaysia, so we were able to collect a rich data set of them. The result shows the effects of the indicators to eachother (figure 3). Prediction results show very accurate modeling .

For the third step, we have applied the methodology for all of the stocks which belongs to FBM100. The result shows that the underlying learned structure is not a directed acyclic graph, which is an essential feature of Bayesian network.
Since the inter-relationships between variables are very complex, it is impossible for us to change the learned model so that it meets DAG structure. Consequently, we have tried to use Markov blanket of each node to predict each node separately. But, we have experienced that when the inter-connection between variables are a lot (it means that the underlying graph is dense), due to the limitations which exist in computation in constraint based algorithms, the learned structure may not be accurate; and cause the prediction to be inaccurate as well. However, these claims do not mean that the learned model is completely incorrect. We have examined some of the nodes for finding out whether the direct links to them are correct or not. For testing this assumption, we have chosen Genting stock as class variable and collected all directly connected nodes to it (figure 4). A summary of the achieved results is tabled below in table 1.

Stock Name	Sign prediction	Amount of change prediction
Construction	92%	52%
finance	92%	80%
Industrial_pr	92%	50%
properties	92%	72%
plantation	78%	66%
Trading and services	96%	39%
Technology	30%	35%
Consumer products	80%	35%
Genting	56.66%	35%

Table 1- total result of using proposed methodology

V. BAYESIAN NETWORK LIMITATIONS

We have found that the learning process will lead to more accurate result if the numbers of variables are few (small dimension) and the numbers of rows (records) are many (big sample size). If the numbers of variables are many and the numbers of rows are few, the learned model will not be accurate and even lead to wrong model.

A limitation in constrained based learning is that the numbers of edges which may be connected to a variable (number of relationships for a variable). It cannot be more than a threshold value, which depends on the computing power of the system.

Another limitation which affects the accuracy of the learned network is the existence of unrelated variable(s) between variable domains. These variables should be avoided and eliminated from variable domain before structure learning takes place. They increase the computational complexity and may lead to the increase of the learning procedure time. In spite of that, these kinds of variables may affect the whole network accuracy and may lead to wrong modelling.

VI. CONCLUTIONS AND FUTURE WORKS

The aim of this study was to introduce learning Bayesian network from stock data as an alternative method for other AI methods and machine learning techniques for stock analysis. We have proposed a methodology for learning Bayesian network form stock data and used the learned model for prognosis and diagnosis against real stocks data. The results show that this approach can be more efficient and straightforward if the structure learning part be improved.

In comparison with other machine learning techniques, Bayesian network is simpler. We have performed our experiment with only one indicator as input and the result was convincing and reasonably good.

The methodology should be applied to other stock market data which are in hour by hour format. Since learned Bayesian network can be more accurate if the data records are more, hour by hour data can be helpful to increase the accuracy of the model.

Feature selection technique should be used for dimension reduction in this area and the technique itself should be improved.

We propose to incorporate knowledge experts' opinions in structure learning. This is needed since the algorithms are not mature to face with large number of variables and very dense graphs (i.e. most of the nodes are inter-related).

Moreover, the model can be embedded into an application program for testing against real time data.

REFERENCES

[1] T. Koski, J. M. Noble. Bayesian networks an introduction, USA: John Wiley,2009,pp1-191. DOI: 10.1002/9780470684023

[2] G. Görz, "Knowledge Representation and Reasoning." Internet:http://www8.informatik.uni-erlangen.de/IMMD8/Lectures/KRR/, May, 14, 2007 [Jun,15,2010].

[3] D. Heckerman, "A tutorial on learning with Bayesian networks". USA: Microsoft Research.1995

[4] J. pearl, "Bayesian Networks". Los Angeles: University of California, Technical Report R-246 , 1997, pp.1-20.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* San Francisco CA:Morgan Kaufman Publishers,1998,pp.1-400

[6] K. Korb, A. Nicholson, *Bayesian artificial intelligence.* London: Chapman & Hall.2003, pp. 1-300.

[7] C. E. Kahn, L. M. Roberts, K. A. Shaffer, P. Haddawy,. "Construction of a Bayesian network for mammographic diagnosis of breast cancer". *Computers in Biology and Medicine* , Vol. 27(1), 1997,pp.19-29.

[8] C. R. Twardy, A. E. Nicholson, K. B. Korb, J. McNeil,(2006) Epidemiological data mining of cardiovascular Bayesian networks. electronic *Journal of Health Informatics* . Vol.1(1),2006, pp. , Available: http://www.csse.monash.edu.au/courseware/cse458/2006/EJHITwardyEtAl2006.pdf [2006]

[9] K. Murphy,"A Brief Introduction to Graphical Models and Bayesian Networks". Internet: http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html, 1998, [03 15, 2010]

[10] M. L.Wong, K. S. Leung, "An Efficient Data Mining Method for Learning Bayesian Networks Using an Evolutionary Algorithm-Based Hybrid Approach". *IEEE Transactions on Evolutionary Computation* ,2004 ,pp.378-404.

[11] J. Cheng, R. Greiner, J. Kelly, D. Bell, W. Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* , Vol.137(1-2), 2002, pp. 43-90.

[12] P. Spirtes, C. Glymour, R. Scheines, Causation Prediction and Search. united states: MIT press, 2000, pp:1-400.

[13] S. Russell, P. Norvig,. artificial inteligence a modern approach. New Jersey: Prentice Hall,2003,pp. 1-500

[14] H. Bengtsson, Bayesian networks-a self-contained introduction with implementation remarks. Sweden: Lund Institute of Technology.2001

[15] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B* , Vol. 50(2), 1998, pp.157-224.

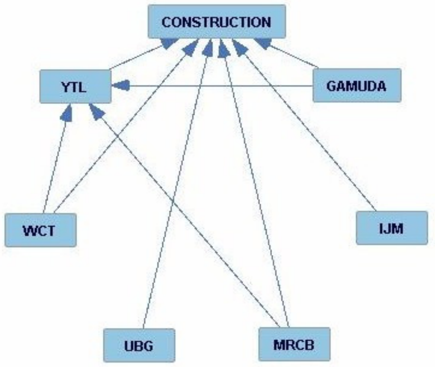

Figure 2. an example of learned bayesian network whithin Comstruction sector

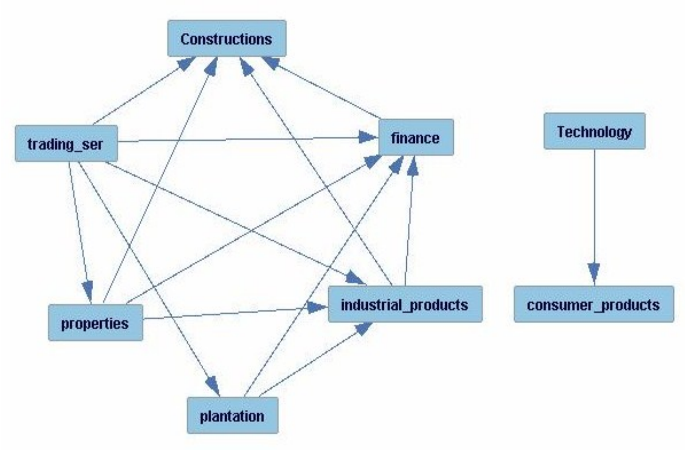

Figure 3. learned Bayesian network between sectors

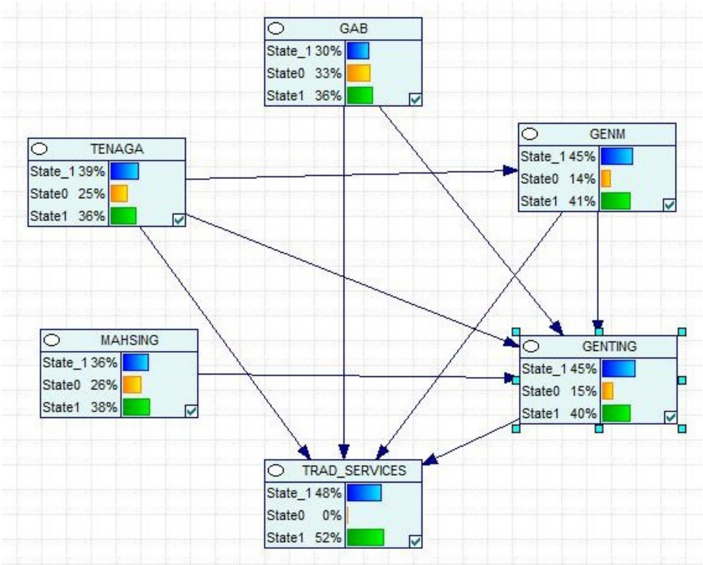

Figure 4. learned Bayesian network for Genting stock prediction