Soft Comput (2016) 20:4575–4588
DOI 10.1007/s00500-015-1764-8

METHODOLOGIES AND APPLICATION

Multiple-cause discovery combined with structure learning for high-dimensional discrete data and application to stock prediction

Weiqi Chen[1] · **Zhifeng Hao**[2] · **Ruichu Cai**[2] · **Xiangzhou Zhang**[3,4] · **Yong Hu**[3,4] · **Mei Liu**[4,5]

Published online: 8 July 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Causal discovery in observational data is crucial to a variety of scientific and business research. Although many causal discovery algorithms have been proposed in recent decades, none of them is effective enough in dealing with high-dimensional discrete data. The main challenge is the complex interactions among large volume of variables, leading to numerous spurious causalities found. In this work, we propose a novel multiple-cause discovery method combined with structure learning (McDSL) to eliminate the spurious causalities. The method is carried out in two phases. In the first phase, conditional independence test is used to distinguish direct causal candidates from the indirect ones. In the second phase, causal direction of multi-cause structure is carefully determined with a hybrid causal discovery method. Validation experiments on synthetic data showed that McDSL is reliable in discovering multi-cause structures and eliminating indirect causes. We then applied this algorithm in discovering multiple causes of stock return based on 13-year historical financial data of the Shanghai Stock Exchanges of China, and established a stock prediction model. Experimental results showed that the McDSL discovered causes revealed changes of key risk factors of the stock market over 13 years, which indicated investors should change their investment strategy over time. Moreover, the causes discovered by McDSL have better performance in predicting stock return than that of other common filter-based feature selection algorithms.

Keywords Causal discovery · High-dimensional discrete data · Structure learning · Additive noise model · Stock prediction

Communicated by V. Loia.

✉ Weiqi Chen
isscwqcxz@126.com

Zhifeng Hao
mazfhao@scut.edu.cn

Ruichu Cai
cairuichu@gmail.com

Xiangzhou Zhang
zhxzhou@mail2.sysu.edu.cn

Yong Hu
henryhu200211@163.com

Mei Liu
meiliu@kumc.edu

1 Faculty of Automation, Guangdong University of Technology, Guangzhou, China

2 Department of Computer Science, Guangdong University of Technology, Guangzhou, China

3 School of Business, Sun Yat-sen University, Guangzhou, China

4 Big Data Decision Institute, Jinan University, Guangzhou, China

5 University of Kansas Medical Center, Kansas City, USA

1 Introduction

Accurate knowledge discovery is a key focus area in Big Data research (Esposito et al. 2015). Causal discovery is a crucial approach for knowledge discovery in many scientific and business research, for example, discovering causal genes of diseases are the focus of biology and medicine (Agbabiaka et al. 2008), retail travel agencies are concerned with the impact of adopting e-business in their supplier relationships (Andreu et al. 2010) and identifying causal factors of stock price fluctuation are of great interests to the investors (Zunino et al. 2010; Zhang et al. 2014).

The recently developed causal discovery methods offer a feasible and economical solution for identifying causalities in observational data, without expensive randomized experiments and interventional experiments. Among them, structure learning (Spirtes et al. 2000; Sobel 1996) and additive noise models (ANMs) (Kano and Shimizu 2003; Mooij et al. 2009) are two mainstream approaches. The structure learning approach is a subfield of Bayesian network learning and focuses on discovering relations among variables; however, it only identifies the Markov equivalent class, and cannot determine the direction of each edge. In contrast to the structure learning approach, ANMs have been proposed to infer direction of a cause–effect pair, but do not work in high-dimensional causal inference problems.

The main difficulties of causal inference in high-dimensional data are due to the complex interactions among its huge number of variables. For instance, given the following example of 5 variables x_1, x_2, x_3, x_4, x_5 and their relationships depicted using $\rightarrow$,

$$x_1, x_2, x_3 \rightarrow x_4 \rightarrow x_5.$$

The variables x_1, x_2, x_3, x_4 may be inferred as the causes of variable x_5 because the causal influence may propagate in the causal structure. Such phenomena are very common in high-dimensional data, and the true causal relations may be buried by the large amount of spurious causalities. In addition, multiple-cause structures also present a great challenge for existing causal discovery method. Existing ANMs for discrete data only work on pairs of variables, and do not consider the complex causal structure. Multiple-cause structure here refers to local causal structures in which a variable has more than 1 cause. In the above example, the 4 variables $\{x_1, x_2, x_3, x_4\}$ form a multiple-cause structure. Both indirect causalities and multiple-cause structures are common in real applications, for example, stock return is affected by several factors at one time (e.g., earning–price ratio, return on equity and amplitude, etc.) (Fama and French 1992; Sethi 1996), and each factor can be influenced by other factors. Therefore, discovering causalities in high-dimensional discrete data is an important effort towards both theoretical investigation and practical application.

In this work, to address the above-mentioned challenges for discovering causalities in high-dimensional discrete data, we propose a novel algorithm called McDSL by taking advantages of both structure learning approach and ANMs. This paper (1) introduces a new algorithm for causal discovery on high-dimensional discrete data, and (2) constructs a stock causal discovery model based on the algorithm. Compared to other causal discovery algorithms, TPCDM has following advantages: (1) reliability—our algorithm is able to discover causalities on various high-dimensional synthetic discrete data with satisfactory recall and precision, and (2) efficiency—the final model can discover multiple causes of stock trend, and more accurately predict stock return using the discovered causes than other modeling algorithms.

This study has two important contributions. First, it proposes a reliable method for discovering multiple-cause structures and distinguishing indirect causes. Both types of causal structures are well known to be hard to solve together by traditional causal discovery algorithms. Second, it provides a stock prediction model based on the proposed causal discovery algorithm using data from the Shanghai Stock Exchanges of China. The model was demonstrated to be both effective in discovering causes of stock trend and yielding good stock return. The varied causes discovered in each training set of stock data indicates that the investors should note the changes and adjust their investment strategy over time. Moreover, the discovered causes of TPCDM performed better than other feature selection algorithms combined with 7 different predictive models for predicting stock return.

The rest of this paper is organized as follows: Sect. 2 describes the related work. Details of the proposed method are given in Sect. 3. Experimental results are presented and analyzed in Sect. 4. Conclusions are given in Sect. 5.

2 Related work

Structure learning and additive noise model are two main approaches for causal discovery on observational data, based on different assumptions of causalities.

2.1 Structure learning

Structure learning is a subfield of Bayesian network learning and is to discover the Markov equivalent classes by applying conditional independent test on the joint distributions (Koller and Sahami 1996; Pearl 2000; Cai et al. 2011). Causal Markov condition and causal faithfulness condition are the basis of this work. Given variable set X, for each variable $x_i \in X$, the Markov blanket (MB) of x_i is a set of variables which includes the parents, children and spouses variables of x_i. The causal Markov condition states that variables are independent of its non-effects given its Markov blanket. Thus, in Fig. 1, the variables $\{x_6, x_7\}$ are independent of x_4 given its Markov blanket $\{x_1, x_2, x_3, x_5\}$. The causal faithfulness condition states that the true probability distribution is faithful to the true causal structure when the true causal structure does not entail a conditional independence relation according to the causal Markov condition, then the conditional independence relation does not hold of the true probability distribution (Cai et al. 2013a; Zhang and Spirtes 2008). Many scholars have employed the structure learning approach for discovering causal relationships and distinguishing the indirect causali-

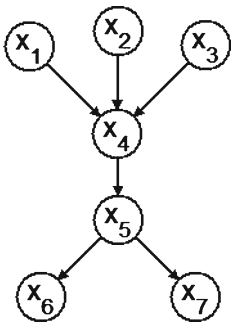

Fig. 1 A running example with 7 variables

ties. Scheines proposed causal Markov condition to discover direct causal variables, and the Markov blanket includes the parent–child–spouse variables of the label variable (Sobel 1996). The MBOR discovered MB under faithfulness condition in large-scale data (De Morais and Aussem 2010). The MAX–MIN strategy was presented to reduce the computation complexity of dependency inference (Tsamardinos et al. 2003). The Markov blanket-embedded genetic algorithm was proposed to discover MB with genetic algorithm (Zhu et al. 2007). In BASSUM, the MB is generated by an iterative growth and pruning process with D-separation strategy (Cai et al. 2011). Tsamardinos et al. described an algorithm for learning local causal structure around target variable to reduce computational complexity (Aliferis et al. 2010). The causal faithfulness condition is employed to discover the spouse and parents with conditional independent test (Cai et al. 2013a; Zhang and Spirtes 2008). Chang et al. (2015) proposed a semantic frame-based topic detection to detect the topic of a document. Karahoca and Tunga (2015) using high-dimensional model representation for constructing a general polynomial model for detecting embolism. Fu et al. (2015) proposed multi-latent Dirichlet allocation models for constructing the categorization system and classify documents. Fernandez-Lozano et al. (2015) employed three feature selection algorithms for classification in a biomedical image texture data set. Although these algorithms are effective in discovering indirect causal variables, they could not infer the causal direction between variables.

D-separation is a frequently used tool to discover the indirect causes and effects of variables from its Markov blanket. The D-separation can be defined as follows: a variable x is independent of its indirect causal variables conditional on at least one subset of its direct causal variables. In a directed acyclic graph, two variables x and y are D-separated, if and only if $\exists s$ that $x \perp y|s$ where s is a set of variables. Pearl (2000) described D-separation as a relation between three distinct variables in a directed acyclic graph (DAG), which is effective in distinguishing direct and indirect causalities in a causal network, but it is unable to infer causal directions.

2.2 Direction learning

The ANM is a recently proposed causal discovery algorithm to determine cause and effect relation of two variables: if causality exists in one direction but not in the reverse, then it can be inferred as the causal direction (Kano and Shimizu 2003; Shimizu et al. 2006). ANMs infer causality between two variables x and y with the distribution of additive noise. More specifically, the joint distribution $P(x, y)$ recognizes the model in one direction, e.g.,

$$y = f(x) + n, \quad n \perp x,$$

But it does not recognize the reversed model, e.g.,

$$x = f(y) + n', \quad n' \perp y.$$

Here, both n and n' are noise. The former causal direction is inferred as the causality between variables x and y, denoted as $x \rightarrow y$. The ANM can infer causality between two variables uniquely and accurately. However, it is difficult for ANMs to infer indirect causal structure like $x \rightarrow y \rightarrow z$, in which causality $x \rightarrow z$ may be inferred incorrectly. Peters et al. (2009) employed the concept of ANMs to autoregressive-moving average (ARMA) for discovering the direction of causal time series. Mooij et al. (2009) used empirical Hilbert–Schmidt independence criterion (HSIC) estimator as the dependence measure in causal direction inference process. Peters et al. (2010, 2011) proved that ANMs is effective on discrete data causal relationship discovery. Hoyer et al. (2009) has shown that the nonlinear models can be solved as the linear models. Cai et al. (2013b) proposed a general Split-and-Merge strategy to discover causality in which the sample size is significantly smaller than the number of variables. However, the researchers only demonstrated that ANMs is effective on low-dimensional data (<8), and it is ineffective when the noise is too small or large than causes. For the multiple-cause structure in Fig. 1, the causes x_2 and x_3 will become extra noise that influences causal algorithms when ANMs infer the causalities between x_1 and x_4.

To discover the causalities on high-dimensional discrete data, the proposed McDSL employs an improved structure learning strategy for distinguishing indirect variables and proposes a causal discovery method for multiple-cause structures by ANMs.

3 McDSL algorithm

The proposed McDSL algorithm for discovering causalities on high-dimensional discrete data includes two phases: (1) structure learning phase (SLP) and (2) direction learning

Algorithm 1 Process of McDSL for discovering causalities in high-dimensional discrete data.

Require: data set $X = \{x_1, \ldots, x_m\}$, variable threshold k.
Ensure: P: The causal skeleton of X.
for $i = 1$ to m **do**
  /*structure learning phase*/
  for each variable $x_j \in X - \{x_i\}$ **do**
    if variable x_j is inferred as a direct causal variables of x_i **then**
      Update C_i with x_j;
    end if
  end for
  /*direction learning phase*/
  for each subset $S \subset \mathfrak{S}$, $\mathfrak{S} = \{S \in 2^{C_i} \mid |S| \leqslant k\}$ **do**
    if $\exists S \in \mathfrak{S}$ that $f(S) \to x_i$ **then**
      Add causalities $\{\forall x_j \in S | x_j \to x_i\}$ into P;
    end if
  end for
end for

phase (DLP). In SLP, the conditional independence relation is exploited to distinguish and exclude the indirect causal variables. The reliable conditional independence test will eliminate spurious causes and dramatically decrease the dimension of candidates for further causal direction inference. The direct causes of variable are denoted as candidates C. In DLP, the causal directions of effect and cause in C are inferred with a hybrid algorithms which includes conditional probability table (CPT) and ANMs.

Algorithm 1 depicts the framework of McDSL on high-dimensional discrete data X with m variables. We aim to discover the causal skeleton P by separately inferring cause(s) of each variable in data X through two phases. For each variable $x_i \in X$, the candidates C_i are discovered from $X - \{x_i\}$ by structure learning. The causes of variable x_i are discovered and added into candidate, and the scale of candidates $|C_i|$ is much smaller than the dimension of entire data $|X|$. Therefore, both accuracy and computational complexity of the causal discovery process are improved in the direction learning phase. If the candidates $C_i \neq \emptyset$, then x_i probability has direct causal variables and its causes will be inferred in the direction learning phase. The direction learning phase includes two stages. First, if $\exists x_j \in C_i$ and the causality $x_j \to x_i$ was inferred, the variable x_j is denoted as its single cause and the causal discovery process of x_i is terminated. Otherwise, x_i may have multiple causes and the second stage is performed to infer the causality between partial subset of its candidates C_i and it. Set S is the partial subset of C_i, and the number of variables in $\forall S \in \mathfrak{S}$ is less than the threshold k. If $\exists S \in \mathfrak{S}$ and the causality $f(S) \to x_i$ had inferred, that the variables $x_j \in S$ are inferred as multiple causes of x_i then the causal discovery process of x_i is terminated. After the causal discovery processes of all variables in X are finished, the entire causality set P is composed of the causalities of each variable and its cause(s) in high-dimensional sparse discrete data.

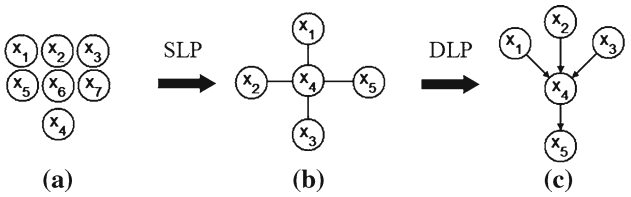

Fig. 2 The causal discovery process of McDSL

Figure 2 shows the process for inferring the causes of x_4 on the data presented in Fig. 2. First, in SLP, the direct causal variables x_1, x_2, x_3, x_5 of x_4 are discovered and denoted as its candidates C_4 in Fig. 2b. Second, in DLP, the causal directions between x_4 and its candidates are inferred. Figure 3 shows the framework of McDSL for discovering causalities on high-dimensional discrete data. We specifically describe the structure learning phase and the direction learning phase in the following subsection.

3.1 Structure learning phase

In high-dimensional discrete data, numerous variables can be indirect causes of a label variable, and this will increase the redundancy and computational complexity for causal discovery. It is extremely challenging for traditional causal discovery methods to eliminate the indirect causal variables. In this study, we employ an indirect variable discrimination method in our algorithm to select the candidates of each variable in high-dimensional discrete data with two stages: growth stage and refinement stage.

3.1.1 The growth stage

The growth stage is employed for discovering potential causes/effects of each variable in high-dimensional sparse discrete data to improve the computational complexity of direction learning phase.

Algorithm 2 shows the process of growth stage for discovering candidates of each variable x_i in m-dimensional discrete data. At the first part of the growth stage, we employed a statistic method based on Chi-square test for calculating the correlations $g(x_i, x_j)$ between variable $x_i \in X$ and $x_j \in X - \{x_i\}$, and the correlations of variable x_i is sorted in descending order which is denoted as X_i'. The variables X_i' (1) are more likely to have the potential causalities with variables x_i, and it is defaulted as the initial candidates C_i. Set 2^{C_i} is the power set of candidates C_i, and S is a partial set of the 2^{C_i} in which the scale of each $S \in \mathfrak{S}$ is less than or equal to the minimal variable between the scale of candidates $|C_i|$ and the variable threshold k. The remaining variables $x_j \in X_i'$ are potentially direct causal variables of x_i and added into C_i, if and only if they are not D-separated by any set of variables $S \in \mathfrak{S}$. When there exists $\exists S \in \mathfrak{S}$ such

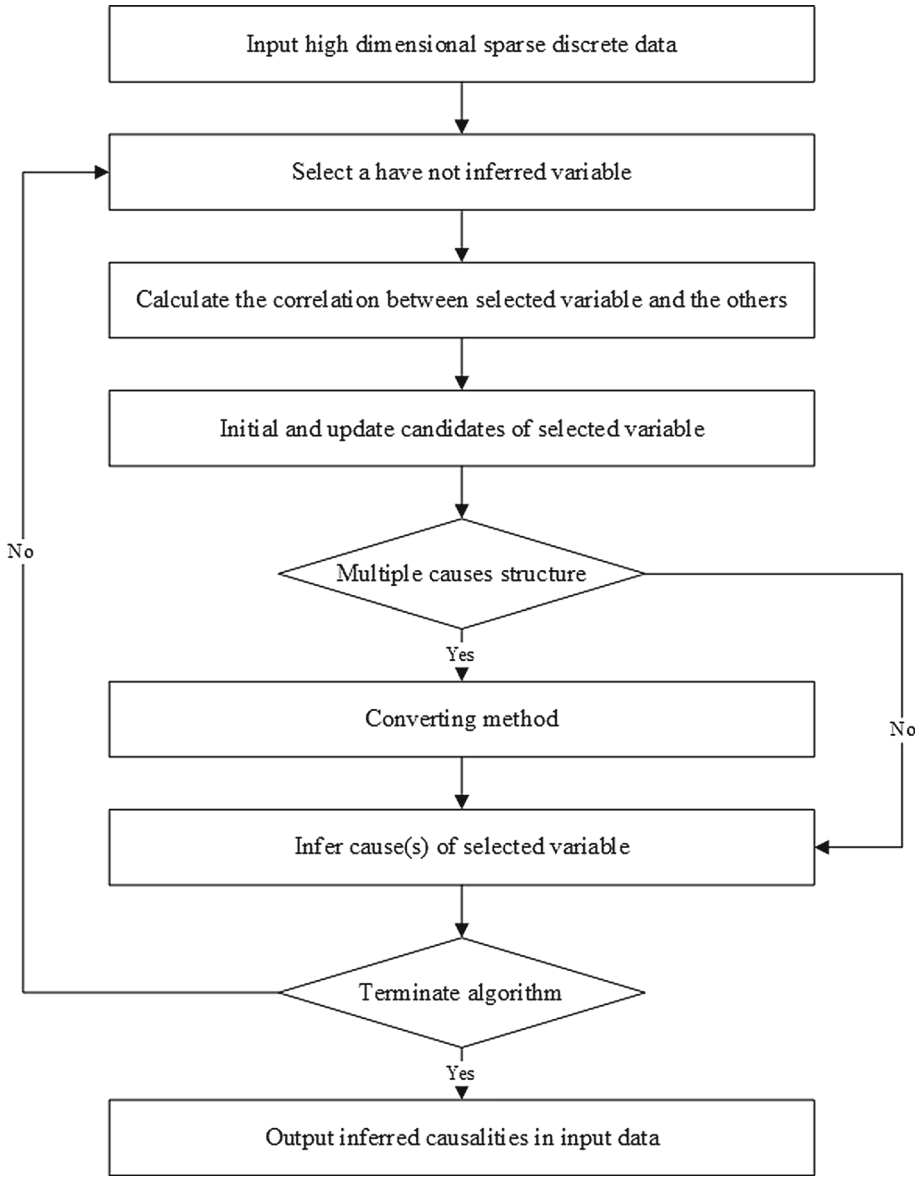

Fig. 3 The framework of McDSL

that $x_i \perp x_j \mid S$, then x_j is inferred as indirect causal variable of x_i and is not added into C_i. Therefore, our algorithm will be able to distinguish most of the indirect causal variable in DLP, and improves accuracy of the causal discovery process.

Accompanied by growing scale of candidates, the computational complexity of growth stage is exponentially increasing. Thus, we propose a pruning method to filter the indirect causal variables from candidates in growth stage to limit the increment in computational complexity. When variable x_j is inferred as a potential direct causal variable of x_i, then the variable $x_k \in C_i$ will be inferred as the indirect causal variable and removed from candidates if and only if x_i and x_k are D-separated by x_j. The pruning method is an effective strategy to infer and remove indirect causal variable from candidates.

3.1.2 The refinement stage

In the growth stage, variables are inferred as potential causes/effects and added into candidates with descending order of correlation. It is possible that the candidates may include few indirect causal variables not filtered out by the pruning method.

In high-dimensional discrete data, variable x_j may indirectly causes variable x_i through multiple paths. Such as the causalities $x_j \rightarrow x_k \rightarrow x_i$ and $x_j \rightarrow x_l \rightarrow x_i$, where variable x_i has two direct causes x_k and x_l, and both of these causes have the same upstream cause x_j. Neither variable x_k nor x_l can distinguish x_j and x_i separately by the pruning method in the growth stage. Therefore, after the growth stage, the refinement stage is to further filter each variable $x_j \in C_i$ and remove indirect causal variables in it.

Algorithm 2 The process of growth stage.

Require: data set $X = \{x_1, ?, x_m\}$, variable threshold k.
Ensure: variable set $C = \{C_1, \cdots, C_m\}$: C_i is the set of potential direct causes/effects of each variable $x_i \in X$.
for $i = 1$ to m **do**
  for each variable $x_j \in X - \{x_i\}$ **do**
    $g(x_i, x_j)$ is correlation between x_i and x_j;
  end for
  Set $X_i' = X - \{x_i\}$, and all variable $x_j \in X_i'$ is sorted in descending order by $G = \{g(x_i, x_j) | x_j \in X - \{x_i\}\}$;
  Add $X_i'(1)$ into C_i
  Remove variable $X_i'(1)$ from X_i';
  for each variable $x_j \in X_i'$ **do**
    Set $\mathfrak{S} = \{S \in 2^{C_i} \mid |S| \leq \min(|C_i|, k|)\}$;
    if $\neq \exists S \in \mathfrak{S}$ that $x_j \perp x_i \mid S$ **then**
      /*pruning method*/
      for each variable $x_k \in \{C_i\}$ **do**
        if $x_k \perp x_i \mid x_j$ **then**
          Remove x_k from C_i;
        end if
      end for
      Add x_j into C_i.
    end if
  end for
end for

Algorithm 3 The process of refinement stage.

Require: variable set $C = \{C_1, \cdots, C_m\}$, variable threshold k.
Ensure: variable set C.
for $i = 1$ to m **do**
  for each variable $x_j \in C_i$ **do**
    $\mathfrak{S} = \{S \in 2^{C_i} \mid |S| \leq \min(|C_i|, k)\}$;
    if $\exists S \in \mathfrak{S}$ that $x_j \perp x_i \mid S$ **then**
      Remove x_j from C_i;
    end if
  end for
end for

Algorithm 3 shows the process of refinement stage for filtering indirect variable in C_i. Set 2^{C_i} is the power set of candidates C_i, and S is a partial set of the 2^{C_i} in which the scale of each $S \in \mathfrak{S}$ is less than or equal to the minimal variable between the scale of candidates $|C_i|$ and the variable threshold k. The variable $x_j \in C_i$ is an indirect cause of x_i if and only if they are D-separated by any subset set $S \in \mathfrak{S}$. When variable x_j is determined as the indirect cause, it will be removed from C_i and the causality of x_j and x_i will not be inferred in DLP.

3.2 Direction learning phase

In high-dimensional discrete data, it is not possible to discover causalities with multiple causes using traditional causal discovery algorithms. In DLP, we propose a hybrid method for inferring the causalities between single variable and a set of variables.

After the structure learning phase, most variables in candidates C_i are potential direct cause(s)/effect(s) of x_i. Then,

Algorithm 4 The process of direction learning phase.

Require: variable set $C = \{C_1, \cdots, C_m\}$, variable threshold k.
Ensure: causal skeleton P.
for $i = 1$ to m **do**
  Set $\mathfrak{S} = \{S \in 2^{C_i} \mid |S| \leq \min(|C_i|, k)\}$;
  for each set $S \in \mathfrak{S}$ **do**
    Set $\overline{x_S} = f(S)$;
    if $\overline{x_S} \rightarrow x_i$ **then**
      Terminate causal discovery process of variable x_i, and add causalities $\{x_j \in S \mid x_j \rightarrow x_i\}$ into P;
    end if
  end for
end for

Table 1 The converting method of set S

	$S(1)$	$\cdots$	$S(q)$	$\overline{x_S}$
States	1	$\cdots$	1	1
	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	1	$\cdots$	K_q	$\prod_{j=2}^{q} K_j$
	2	$\cdots$	1	$\prod_{j=2}^{q} K_j + 1$
	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	K_1	$\cdots$	K_q	$\prod_{j=1}^{q} K_j$

we infer the causalities between set of variables S and x_i to discover true cause(s) with a converting method. Algorithm 4 shows the process of direction learning phase. Set 2^{C_i} is the power set of candidates C_i, and S is a part of the 2^{C_i} in which the scale of each $S \in \mathfrak{S}$ is less than or equal to the minimal variable between the scale of candidates $|C_i|$ and the variable threshold k. Our algorithm converts the multiple-cause structure $S \rightarrow x_i$ into single-cause structure $f(S) \rightarrow x_i$ using conditional probability table, that those two causal structures are equal. For instance, the set of variables $S = \{S(1), \ldots, S(q)\}$ is a subset of candidates C_i, and each variable $S(j) \in S$ has K_j states, as shown in Eq. (1).

$$\begin{cases} S = \{S(1), \ldots, S(q)\}, & S \subset C_i \\ S(j) = \{1, \ldots, K_j\}, & j = 1, \ldots, q \end{cases} \tag{1}$$

The converting process of $\overline{x_S} = f(S)$ is given in Table 1.

As shown in Table 1, the distribution of states of converted variable $\overline{x_S}$ is influenced by all variables in set S. For two different sets of variables $S, S' \subset C_i$, the converted variables $\overline{x_S}$ and $\overline{x_{S'}}$ are not the same. Therefore, we propose the following corollary.

Corollary 1 *Set* $\forall S, S' \in \mathfrak{S}$, $f(S) = \overline{x_S}$ *and* $f(S') = \overline{x_{S'}}$. *If* $S \neq S'$, *then* $\overline{x_S} \neq \overline{x_{S'}}$.

From Corollary 1, different sets of variables S correspond to different converted variables $\overline{x_S}$. Therefore, the causalities between $\overline{x_S}$ and x_i can be inferred when S is the set of its all causes.

Corollary 2 *Set $S \in \mathfrak{S}$ is subset of candidates C_i and $f(S) = \overline{x_S}$. Then, the converted variable $\overline{x_S}$ is the cause of x_i if and only if S is the set of all causes of x_i.*

From Corollary 1 and Corollary 2, we propose following definition for inferring causalities between set of variables and single variable by converting method.

Definition 1 Set C_i is candidates of variables, 2^{C_i} is power set of C_i which is denoted as $2^{C_i} = \{S | S \subset C_i\}$, and $S = \{S \in 2^{C_i} \mid |S| \leq \min(|C_i|, k)\}$. If $\exists S \in \mathfrak{S}$, that $f(S) = \overline{S}$, $\overline{x_S} \to x_i$, and $\forall S' \in 2^{C_i}, S \neq S'$ that $f(S') = \overline{x_{S'}}, \overline{x_{S'}} \nrightarrow x_i$. Then, S is a set of causes of x_i, which is denoted as $S \to x_i$.[1]

Based on the above corollary and definition, causalities of variable with multiple causes can be accurately inferred in DLP. The learning phase of $\forall x_i \in X$ will be terminated when the set of true cause(s) has inferred. Due to the effort of SLP, the scale of candidates is limited and reduced. Moreover, the variable threshold k is employed to improve the inferring efficiency both in SLP and DLP. Therefore, our proposed can effectively discover the causalities in high-dimensional sparse discrete data.

4 Experiments and application

In this section, we assess the performance of our algorithm on both synthetic and stock market data sets to investigate whether the algorithm is able to rediscover true multiple-cause structures effectively. Section 4.1 simulates several discrete synthetic data and divides them into 4 types of experiments to evaluate the accuracy of our algorithm. Section 4.1.1 presents the performance of our algorithm on causal structures with 2–9 causes. Section 4.1.2 supplements the theoretical results with randomly generated low-dimensional discrete data, and all those data have double-cause structures and indirect causalities. The Sect. 4.1.3 shows how McDSL performs in discovering the entire causalities of synthetic sparse discrete data sets using different samples and variables. In addition to the synthetic data sets, in Sect. 4.2, we employed McDSL on stock market data obtained from the Shanghai Stock Exchanges as an application. Section 4.2.1 shows the discovered causes of stock return, and most causes are known. Section 4.2.2 discusses the stock trend prediction of McDSL with the discovered causes. The experiments result demonstrates that McDSL performs well compared to other algorithms, and it can provide an intelligent decision support tool for investors to effectively predict stock trend through causal discovery.

[1] If $|S|, |S'| = 1$, that the above definition will be transformed into the definition in article (Peters et al. 2011).

4.1 Experiments on synthetic data

4.1.1 Discovering multiple-cause structures

In this paper, we employed the model that Peters et al. (2011) proposed, and extended it to design synthetic multiple-cause structures for investigating the performance of McDSL.

$$y = \begin{cases} x + N, & \text{if } \ x \to y \\ f(x_1, \ldots, x_m) + N, & \text{if } \ x_1, \ldots, x_m \to y \end{cases} \tag{2}$$

The designed synthetic model as shown in Eq. (2),variable $x_i \sim B(2, p)$ is the cause of variable y, and variable $N \sim B(2, p)$ is noise, parameter $p \in [0.1, 0.9]$. The function $f(x_1, \ldots, x_m)$ is the convening process of m variables by conditional probability table, supp $f(x_j, \ldots, x_m) \in \{0, 1\}$ and supp $x_i \in \{0, 1, 2\}$. We had generated 300 difference causal structures, and our algorithm is executed with a significance level $\alpha = 0.05$ for the independent test, and the variable threshold $k = 3$.

Peters et al. (2011) proposed 4 evaluation indexes for synthetic experiments: 'Corr dir' indicates the discovered causal direction is equal to the given one; 'Wrong dir' indicates the discovered causal direction is reverse to the given one; 'Both dir' indicates the causalities are discovered in both directions; 'None dir' indicates the causalities cannot be discovered in both directions. We divided this into 2 evaluation indexes to investigate the McDSL performance in synthetic experiments: 'Correct' indicates the discovered causal direction is equal to the given one; 'Wrong' indicates the discovered causal direction is not equal to the given one.

Table 2 shows that McDSL achieved high accuracies on both sparse ($n \in [1, 3]$) and complex causal structures ($n > 3$). McDSL was able to exclude all false causalities since the probabilities are $\leq$0.15. Although the accuracies of McDSL are lower than 90 % when $n \in \{4, 5\}$, the inferred causalities by our algorithm are still reliable.

4.1.2 Distinguishing incorrect and indirect causalities

We design two causal skeletons in Fig. 4 to investigate the performance of McDSL in low-dimension discrete data, and we employed the model in Eq. (2) to generate 300 different synthetic data for each skeleton. Both these causal skeletons had 5 variables and include indirect causalities and multiple-cause structures.

Table 2 The performance of our algorithm in inferring 9 different causal structures

$x_1, \ldots, x_n \rightarrow y$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$
Correct:	0.9600	0.9300	0.9067	0.8733	0.8500	0.9133	0.9567	0.9833
Wrong:	0.0400	0.0700	0.0933	0.1267	0.1500	0.0867	0.0433	0.0167

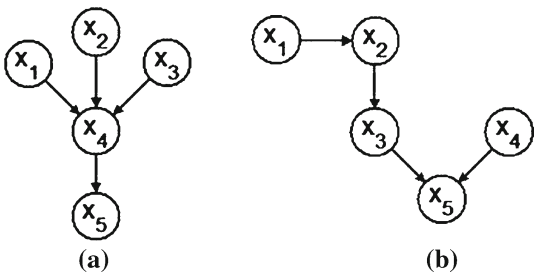

Fig. 4 Two synthetic low-dimensional discrete data

In Tables 3 and 4, the correct multiple-cause structure $x_1, x_2, x_3 \rightarrow x_4$ in Fig. 4a and $x_3, x_4 \rightarrow x_5$ in Fig. 4b can be discovered by our algorithm with high accuracy (0.92 and 0.9633, respectively). The Definition 1 in Sect. 3.2 shows that our algorithm can discover correct causality if and only if the set of causes is complete, and the experimental results in Tables 3 and 4 prove the point. The false causalities are also correctly distinguished in Table 3 (92.00 % in causality $x_1 \rightarrow x_4$, 100.00 % in causality $x_1, x_2 \rightarrow x_4$ and 100.00 % in causality $x_1, x_2, x_3, x_5 \rightarrow x_4$) and Table 4 (96.67 % in causality $x_1, x_4 \rightarrow x_5$ and 96.00 % in causality $x_2, x_4 \rightarrow x_5$), respectively.

4.1.3 Discovering causalities on high-dimensional synthetic discrete data

In this section, we aimed to investigate the performance of our algorithm in synthetic high-dimensional sparse discrete data. Therefore, we have generated 5 different high-dimensional sparse acyclic skeletons (15, 35, 80, 100 and 150 variables) for evaluation. Each skeleton has 40 % double-cause structures and 60 % single-cause structures. We illustrate two skeletons with 15 (a) and 35 (b) variables in Fig. 5 separately.

For each skeleton, we generated 7 data sets (50, 200, 500, 1000, 2000, 10,000 and 20,000 samples) by the model in Eq. (2) to evaluate the influence of sample size. Recall, precision and F value are used as the measure of performance as follows.

Table 4 The accuracy of McDSL for discovering causalities on 300 different discrete data sets in Fig. 4b

	$x_3, x_4 \rightarrow x_5$	$x_1, x_4 \nrightarrow x_5$	$x_2, x_4 \nrightarrow x_5$	$x_1 \nrightarrow x_3$
Correct:	0.9633	0.9667	0.9600	0.9233
Wrong:	0.0367	0.0333	0.0400	0.0767

$$\text{precision} = \frac{(|\{\text{discovered causalities}\} \bigcap \{\text{actual causalities}\}|)}{(|\{\text{discovered causalities}\}|)}$$

$$\text{recall} = \frac{(|\{\text{discovered causalities}\} \bigcap \{\text{actual causalities}\}|)}{(|\{\text{actual causalities}\}|)}$$

$$F\text{value} = \frac{(2 * \text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

Figure 6a investigates the influence of different sample sizes on McDSL and ANM using 100 variables. Neither algorithm is appealing for causal discovery when the sample size is 50 and ANM's performance peaked at a low F value of approximately 0.6 when sample size is more or equal to 200. In contrast to ANM, performance of McDSL increased steadily with the increasing sample size, achieving the highest F value of 92.72 % with 10,000 samples. McDSL performed well with 2000 samples, and since the training sample size for stock data is close to 2000, we set the sample size as 2000 in following evaluation.

Figure 6b examines two algorithms on different dimensions (i.e., different number of variables) using 2000 samples. The recalls of ANM is 0.5 for the 15-variable data and 0.4 for other data, which means that it has merely discovered less than or equal to 50 % causalities in each synthetic data. Nonetheless, the precision of ANM is equal to 1 in all synthetic data, which is evident that ANM is a precise algorithm for causal discovery. On the other hand, McDSL can accurately discover the causal network of each synthetic data, and the precision and recall of each synthetic data both fluctuate around 0.9. For instance, the recall and precision of McDSL for the 15-variable data are both equal to 1, and it achieved 88.29 % in recall and 93.30 % in precision for the 150-variable data. Therefore, McDSL can comprehensively

Table 3 The accuracy of McDSL for discovering causalities on 300 different discrete data sets in Fig. 4a.

	$x_1, x_2, x_3 \rightarrow x_4$	$x_1 \nrightarrow x_4$	$x_1, x_2 \nrightarrow x_4$	$x_1, x_2, x_3, x_5 \nrightarrow x_4$	$x_1, x_2, x_3 \nrightarrow x_5$
Correct:	0.9200	0.9200	1.0000	1.0000	0.8833
Wrong:	0.0800	0.0800	0	0	0.1167

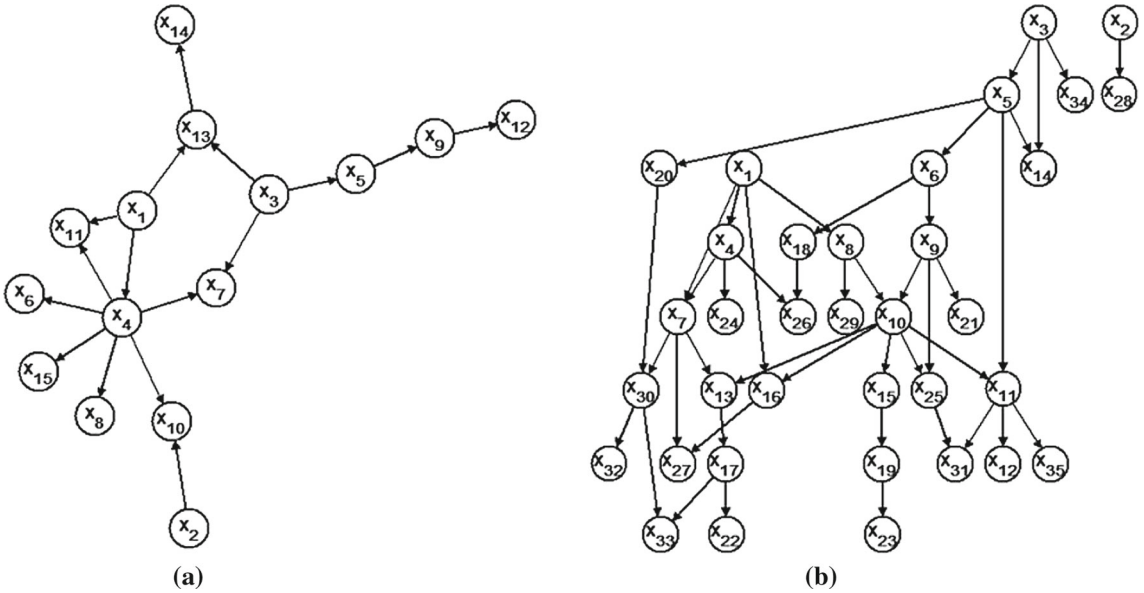

Fig. 5 Synthetic sparse causal skeletons with 15 (**a**) and 35 (**b**) variables

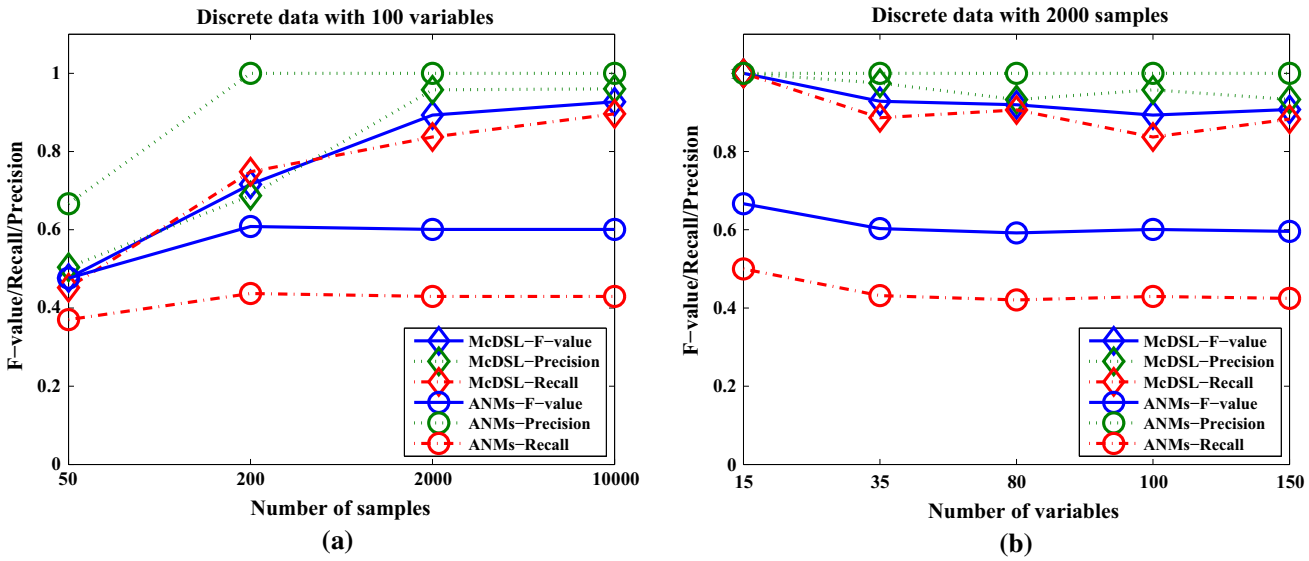

Fig. 6 Comparison of McDSL and ANMs on different samples (**a**) and variables (**b**)

and accurately discover the causalities in the synthetic high-dimensional sparse discrete data with 2000 samples.

As an illustration, Fig. 7 shows the simulated causal network of 35 variables and the discovered causal network by MCDSL with 2000 samples (88.64 % recall, 92.86 % precision). The spurious causalities will possibly prevent our method from accurately discovering the multiple-cause structure. For instance, in Fig. 7a, the variable x_1 not only indirectly causes x_7 in $x_1x_4 \rightarrow x_7$, but also directly causes x_7, so that the variable x_1 is discovered as the cause of variable x_7 in the estimating stage of the discovering phase and then multiple-cause discovering stage is terminated. However, most of the causalities in this simulated data set had been discovered, which demonstrate that our method is reliable in discovering causalities in high-dimensional discrete data.

4.2 Experiments on stock data

For further analysis, our algorithm is applied to stock data, which is previously proposed by Zhang et al. (2014), to select representative factors for stock prediction. The experimental data are obtained from the annual financial report of A-shares of the Shanghai Stock Exchanges and covers the period of 2000 to 2012. The data set has 50 input variables (stock factors) and 1 output variable (stock return). The 50 potential

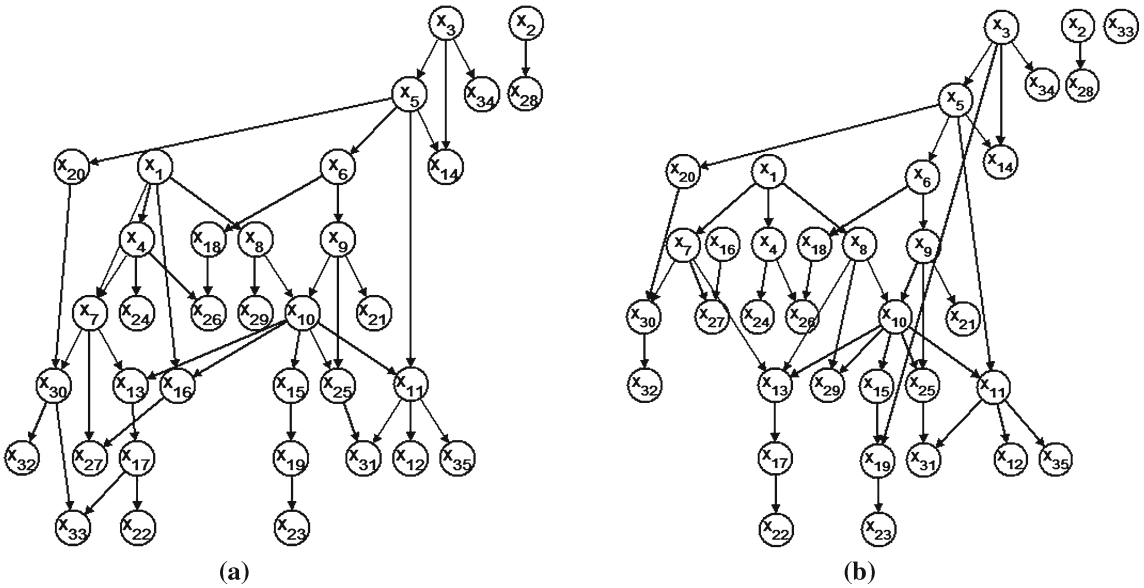

Fig. 7 The actual (**a**) and discovered (**b**) causal networks of 35 variables with 2000 samples

causal factors of stock return are divided into 12 categories: valuation, profitability, growth, leverage, liquidity, operation, momentum, size, cash flow, bonus, volatility and volume. The factors are given by Zhang et al. (2014), and cases with missing data are eliminated to reduce its influence on performance. Zhang et al. (2014) divided the stock data of 13 years into 9 overlapping in-sample estimation periods (2000–2004, 2001–2005,. . .,2008–2012), and each period included a training set (4 years) and a testing set (1 year). In Sect. 4.2.1, we employ McDSL to discover causes(s) of stock tend from 50 potential causal factors in each training set, because the causal factors of stock return might vary in time. Section 4.2.2 uses the discovered cause(s) to predict stock trends of testing set in the same overlapping in-sample estimation periods. Experimental results of both Sects. 4.2.1 and 4.2.2 show McDSL is effective for discovering concept drift in predicting stock trends. Moreover, our algorithm performed better than several classical feature section algorithms.

4.2.1 Discovering causal factors of stock trend

We applied McDSL to discover the cause(s) of the stock return in each training set, with a significance level of $\alpha = 0.05$ for the independent test, and a variable threshold k of 15. The results are presented in Tables 5 and 6.

Table 5[2] shows the result of McDSL for causal discovery on 9 overlapping in-sample estimation training sets separately, and each training set has 6–10 discovered direct causes. There are 26 factors inferred as the direct causal factor for stock return in all training sets, and they are distributed in 10 categories: "Valuation", "Profitability", "Growth", "Leverage", "Operation", "Momentum", "Size", "Cash flow", "Volatility" and "Volume". Moreover, the 10 categories can be divided into 2 groups of indicators: "Fundamental indicators" and "Technical indicators". The "Fundamental indicators" focuses on the enterprise intrinsic value, that it includes enterprise assets and financial expense. The "Technical indicators" focuses on the stock price, that it includes fluctuation of stock market and stock price history. Both the "fundamental indicators" and "technical indicators" were employed to predict the trend of stock price.

To distinguish the importance of discovered causes, we sorted them in descending order of frequency (6), and only factors that were inferred as causes in at least 3 periods were listed. The most frequent 5 factors are: "#1 E/P", "#31 Buy–Hold Return 3-month%", "#33 Buy–Hold Return 12-month%", "#36 Circulation market value to total market value", and "#42 Amplitude 6-month%". These factors are belonged to 4 categories: "valuation", "Momentum", "Size" and "Volatility", respectively.

The distributions of discovered causes have drifted in 9 training sets, and there are obvious difference between the 5th training set (2004–2007) and the 6th training set (2005–2008). This difference deserves further discussion. Those training sets involved two important events in the China stock market, i.e., the great crash in late 2007 and the 2008 Financial Crisis. Satisfyingly, McDSL discovered the concept drift of causes. For instance, (1) in the first 5 training sets (2000–2003,?,2004–2007), factor "#1 E/P" was discovered as the causal factor of stock return 5 times, factors "#4 ROE" and "#5 ROA" were discovered as the causal factor of stock return

[2] # Factor represents that Factor # is inferred as the causes of return in training set by McDSL.

Table 5 The discovered cause(s) of stock return in each training period

Training period	Discovered cause(s)	
2000–2003	#1 E/P	#29 Long-term liabilities to operating capital
	#2 B/P	#42 Amplitude 6-month%
	#15 Total assets%	#43 Amplitude 12-month%
	#21 Current liabilities rate	#44 SD of daily return rate 3 months
2001–2004	#1 E/P	#33 Buy–Hold Return 12-month%
	#4 ROE	#34 Circulation market value
	#5 ROA	#36 Circulation market value to total market value
	#29 Long-term liabilities to operating capital	#45 SD of daily return rate 6 months
	#31 Buy–Hold Return 3-month%	
2002–2005	#1 E/P	#21 Current liabilities rate
	#4 ROE	#35 Total market value
	#5 ROA	#36 Circulation market value to total market value
	#7Net profit margin on sales	#49 Turnover to total market turnover 1-month%
2003–2006	#1 E/P	#31 Buy–Hold Return 3-month%
	#3 S/P	#42 Amplitude 6-month%
	#5 ROA	#43 Amplitude 6-month%
	#15 Total assets%	#45 SD of daily return rate 6 months
	#21 Current liabilities rate	
2004–2007	#1 E/P	#45 SD of daily return rate 6 months
	#4 ROE	#47 Turnover rate 1-month%
	#42Amplitude 6-month%	#48 Turnover rate 3-month%
2005–2008	#26 Inventory turnover	#34 Circulation market value
	#31 Buy–Hold Return 3-month%	#42 Amplitude 6-month%
	#32 Buy–Hold Return 6-month%	#47 Turnover rate 1-month%
	#33 Buy–Hold Return 12-month%	#50 Turnover to total market turnover 3 month%
2006–2009	#4 ROE	#34 Circulation market value
	#16 Leverage%	#35 Total market value
	#31 Buy–Hold Return 3-month%	#36 Circulation market value to total market value
	#32 Buy–Hold Return 6-month%	#49 Turnover to total market turnover 1-month%
	#33 Buy–Hold Return 12-month%	
2007–2010	#15 Total assets%	#36 Circulation market value to total market value
	#31 Buy–Hold Return 3-month%	#40 Operating net cash flow
	#33 Buy–Hold Return 12-month%	#42 Amplitude 6-month%
	#34 Circulation market value	#49 Turnover to total market turnover 1-month%
	#35 Total market value	#50 Turnover to total market turnover 3-month%
2008–2011	#15 Total assets%	#36 Circulation market value to total market value
	#21 Current liabilities rate	#40 Operating net cash flow
	#29 Long-term liabilities to operating capital	#45 SD of daily return rate 6 months
	#33 Buy–Hold Return 12-month%	#49 Turnover to total market turnover 1-month%
	#35 Total market value	#50 Turnover to total market turnover 3-month%

3 times, and all of them are fundamental indicators; (2) in the last 4 training sets (2005–2008, ...,2008–2011), the factor "#49 Turnover to total market turnover 1 month" was discovered as the causal factor of stock return 3 times, factors "#33 Buy-ss-Hold return 12month%" and "#50 Turnover to total market turnover 3 month%" were discovered 4 times, and all of them are technical indicators. Therefore, we have identified an interesting phenomenon that the focus of investors at this point should change from "Fundamental indicators" to "Technical indicators".

Table 6 The frequency of each discovered cause in total 9 training

Factor	Times	Factor	Times
#1 E/P	5	#34 Circulation market value	4
#31 Buy–Hold Return 3-month%	5	#35 Total market value	4
#33 Buy–Hold Return 12-month%	5	#45 SD of daily return rate 6 months	4
#36 Circulation market value to total market value	5	#49 Turnover to total market turnover 1-month%	4
#42 Amplitude 6-month%	5	#50 Turnover to total market turnover 3-month%	4
#4 ROE	4	#5 ROA	3
#15 Total assets%	4	#29 Long-term liabilities to operating capital	3
#21 Current liabilities rate	4		

Table 7 The normalized values of return of different stock trend prediction model

	NoFS	McDSL	CFS	CART	LASSO
LR	1 (+0.13)	0.87	0.96 (+0.09)	0.92 (+0.05)	0.88 (+0.01)
NB	1 (+0.02)	0.98	0.93 (−0.05)	0.96 (−0.02)	0.99 (+0.01)
BN	0.99 (+0.02)	0.97	0.93 (−0.04)	0.97 (0)	1 (+0.03)
NN	1 (+0.17)	0.83	0.99 (+0.16)	0.76 (−0.07)	0.94 (+0.17)
SVM	0.91 (−0.09)	1	0.92 (−0.08)	0.82 (−0.18)	0.87 (−0.13)
J48	0.93 (−0.07)	1	0.94 (−0.06)	0.89 (−0.11)	0.90 (−0.1)
RF	0.89 (−0.09)	0.98	1 (+0.02)	0.85 (−0.13)	0.91 (−0.07)
Average	0.96 (+0.01)	0.95	0.95 (0)	0.88 (−0.07)	0.93 (−0.02)

4.2.2 Predicting stock return

After the discovery of the causal features, we used the same evaluation scheme as adopted in Zhang et al. (2014) to provide a clear performance comparison between McDSL and other feature selection methods in the stock prediction scenario. They employed several popular data mining algorithms to establish distinct stock trend prediction models to predict the stock return, including logistic regression (LR), neural network (NN), support vector machine (SVM), decision tree (DT), Bayesian network (BN) and naïve Bayes (NB) (Tsai et al. 2011; Tsai and Hsiao 2010; Zuo and Kita 2012; Lee 2009), in combination with several feature selection methods, including CFS, CART and LASSO (Zhang et al. 2014; Breiman et al. 1984; Tibshirani 1994). Their methods and results were used as the benchmark in our experiment.

Table 7 [3] presents the investment return of the combination of each feature selection algorithms and stock trend prediction models. The experimental result shows that none of the 4 feature selection methods can produce high investment return using baseline models LR, NB and NN. However, McDSL performed the best when combined with either SVM or J48. Specifically, our algorithm is more reliable in predicting investment return than CART when combined with 6 baseline models. LASSO achieved better investment return than McDSL when combined with LR (1 %), NB (1 %), BN (3 %) and NN (11 %). However, McDSL is better than LASSO when combined with SVM (13 %), J48 (10 %) and RF (7 %). The performance of CFS is close to our algorithm, and they have the same mean value of 0.96. CFS is better than MCDSL with LR (9 %), NN (16 %) and RF (2 %). But McDSL is more effective than the remaining algorithms.

[3] 'NoFS' indicates no feature selection. Best results are highlighted in bold. The value in parentheses indicates the performance difference with the corresponding our algorithm. 'Average' is the average value of 6 algorithms on 7 baseline models.

5 Conclusions and discussion

To perform better causal discovery on high-dimensional discrete data, discovering multi-cause structures and distinguishing indirect causalities are important. Structure learning approach performs well in distinguishing indirect causalities, but it cannot infer causal directions. Conventional causal discovery algorithm, such as ANM is effective in solving single-cause structures, but cannot infer multiple-cause structures or distinguish indirect causalities. Neither structure learning nor ANMs independently work well in discovering causalities from high-dimensional discrete data. This study

proposes a multiple-cause discovery combined with structure learning (McDSL) to address the challenges. The proposed MsDSL method is a completely new approach, suitable for sparse data set. Performance of the proposed algorithm was evaluated on both synthetic and real-world stock data.

The comparative experiments showed that our study can precisely discover causalities from high-dimensional discrete data. In the synthetic data sets, McDSL performed well in both discovering multiple-cause structures and distinguishing indirect and incorrect causalities, and it can effectively and stably discover the entire causal network from high-dimensional discrete data with variable size larger than 100 and sample size larger than 2000. In the discrete stock data, McDSL had discovered several causal factors of stock trend for each training period, and most of the causes were also recognized by other researchers. Moreover, our algorithm was compared to three popular algorithms in predicting stock return, namely, CART, LASSO and CFS. When combined with each of the seven models (i.e., LR, NB, BN, NN, SVN, J48 and RF), McDSL outperformed CART, LASSO and CFS in most cases. It is worth mentioning that our algorithm had the best performance when it is combined SVM and J48.

The most important contribution of this the development of McDSL for causal discovery on high-dimensional discrete data. Moreover, McDSL can discover direct causes (factors) of stock return, and can identify concept drift phenomenon where causes can change over different periods of time. Although McDSL is shown to be an effective algorithm in solving causal discovery problem on high-dimensional discrete data, room still remains for improvement. For instance, the threshold used for the causality discovering phase is complicated to set to achieve balance between the computational complexity and accuracy, in which we had to pick empirically in our experiments. An inappropriate threshold will affect the accuracy of discovering multiple-cause structures, and in future work, we would focus on designing a heuristic algorithm for generating threshold on different high-dimensional data sets.

Acknowledgments This research was partly supported by the National Natural Science Foundation of China (71271061, 70801020), Science and Technology Planning Project of Guangdong Province, China (2010B010600034, 2012B091100192), Guangdong Natural Science Foundation Research Team (S2013030015737), and Business Intelligence Key Team of Guangdong University of Foreign Studies (TD1202).

References

Agbabiaka TB, Savović J, Ernst E (2008) Methods for causality assessment of adverse drug reactions. Drug Saf 310(1):21–37

Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD (2010) Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. J Mach Learn Res 11:171–234

Andreu L, Aldás J, Bigné JE, Mattila AS (2010) An analysis of e-business adoption and its impact on relational quality in travel agency-supplier relationships. Tour Manag 310(6):777–787

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton

Cai R, Zhang Z, Hao Z (2011) Bassum: a Bayesian semi-supervised method for classification feature selection. Pattern Recognit 440(4):811–820

Cai R, Zhang Z, Hao Z (2013a) Causal gene identification using combinatorial v-structure search. Neural Netw 43:63–71

Cai R, Zhang Z, Hao Z (2013b) Sada: a general framework to support robust causation discovery. In: Proceedings of the 30th international conference on machine learning, pp 208–216

Chang YC, Hsieh YL, Chen CC, Hsu WL (2015) A semantic frame-based intelligent agent for topic detection. Soft Comput. doi:10.1007/s00500-015-1695-4

De Morais SR, Aussem A (2010) A novel Markov boundary based feature subset selection algorithm. Neurocomputing 730(4):578–584

Esposito C, Ficco M, Palmieri F, Castiglione A (2015) Smart cloud storage service selection based on fuzzy logic, theory of evidence and game theory. IEEE Trans Comput. doi:10.1109/TC.2015.2389952

Fama EF, French KR (1992) The cross-section of expected stock returns. J Financ 470(2):427–465

Fernandez-Lozano C, Seoane JA, Gestal M, Gaunt TR, Dorado J, Campbell C (2015) Texture classification using feature selection and kernel-based techniques. Soft Comput doi:10.1007/s00500-014-1573-5

Fu R, Qin B, Liu T (2015) Open-categorical text classification based on multi-lda models. Soft Comput 190(1):29–38

Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B (2009) Nonlinear causal discovery with additive noise models. In: Advances in neural information processing systems, pp 689–696

Kano Y, Shimizu S (2003) Causal inference using nonnormality. In: Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion, pp 261–270

Karahoca A, Tunga MA (2015) A polynomial based algorithm for detection of embolism. Soft Comput 190(1):167–177

Koller D, Sahami M (1996) Toward optimal feature selection. Proc int conf mach Learn 20(1113):284–292

Lee M-C (2009) Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Syst Appl 360(8):10896–10904

Mooij J, Janzing D, Peters J, Schölkopf B (2009) Regression by dependence minimization and its application to causal inference in additive noise models. In: Proceedings of the 26th annual international conference on machine learning, pp 745–752. ACM

Pearl J (2000) Causality: models, reasoning and inference, vol 29. Cambridge Univ Press, Cambridge

Peters J, Janzing D, Gretton A, Schölkopf B (2009) Detecting the direction of causal time series. In: Proceedings of the 26th annual international conference on machine learning, pp 801–808. ACM

Peters J, Janzing D, Schölkopf B (2010) Identifying cause and effect on discrete data using additive noise models. In: International conference on artificial intelligence and statistics, pp 597–604

Peters J, Janzing D, Scholkopf B (2011) Causal inference on discrete data using additive noise models. IEEE Trans Pattern Anal Mach Intell 330(12):2436–2450

Sethi R (1996) Endogenous regime switching in speculative markets. Struct Change Econ Dyn 70(1):99–118

Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A (2006) A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res 7:2003–2030

Sobel ME (1996) An introduction to causal inference. Sociol Methods Res 240(3):353–379

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search, vol 81. MIT press, Cambridge

Tibshirani R (1994) Regression shrinkage and selection via the lasso. J Royal Stat Soc 58(1):267–288

Tsai C-F, Hsiao Y-C (2010) Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis Support Syst 500(1):258–269

Tsai C-F, Lin Y-C, Yen DC, Chen Y-M (2011) Predicting stock returns by classifier ensembles. Appl Soft Comput 110(2):2452–2459

Tsamardinos I, Aliferis CF, Statnikov A (2003) Time and sample efficient discovery of markov blankets and direct causal relations. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp 673–678. ACM

Zhang J, Spirtes P (2008) Detection of unfaithfulness and robust causal inference. Minds Mach 180(2):239–271

Zhang X, Yong H, Xie K, Wang S, Ngai EWT, Liu M (2014) A causal feature selection algorithm for stock prediction modeling. Neurocomputing 142:48–59

Zhu Z, Ong Y-S, Dash M (2007) Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognit 400(11):3236–3248

Zunino L, Zanin M, Tabak BM, Pérez DG, Rosso OA (2010) omplexity-entropy causality plane: A useful approach to quantify the stock market inefficiency. Phys A Stat Mech Appl 3890(9):1891–1901

Zuo Y, Kita E (2012) Stock price forecast using Bayesian network. Expert Syst Appl 390(8):6729–6737