Decision Support Systems 106 (2018) 30–43

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Change detection model for sequential cause-and-effect relationships

Tony Cheng-Kui Huang[a], Pu-Tai Yang[b,*], Jen-Hung Teng[c]

[a]Department of Business Administration, National Chung Cheng University, Chiayi, Taiwan, ROC
[b]Department of Business Administration, Tunghai University, Taichung, Taiwan, ROC
[c]Department of Information Management, National Chung Cheng University, Chiayi, Taiwan, ROC

ARTICLE INFO

Article history:
Received 27 July 2016
Received in revised form 31 October 2017
Accepted 27 November 2017
Available online 5 December 2017

Keywords:
Data mining
Change mining
Classifiable sequential patterns
Cause-and-effect relationships
Big data

ABSTRACT

Detecting changes of behaviors or events is crucial when updating existing knowledge in a dynamic business environment. Currently, data analysts can immediately collect data and easily access existing knowledge. However, that knowledge can also rapidly become outdated. This study discusses a form of knowledge, *classifiable sequential patterns* (*CSP*s), defined as $s \rightarrow c$, where s is a temporal sequence; c is a class label; and "$\rightarrow$" is a sign which implies the sequential relationships between s (cause) and c (effect). If the *CSP* evolves into another, and the new knowledge is not updated, decision-makers would continue to work with the obsolete *CSP*. To the authors' knowledge, no study has addressed the topic of change mining in CSPs. To address this research gap, this study proposes a novel change-mining model, *SeqClassChange*, to identify changes in CSPs. Experiments were conducted with a real-world dataset to evaluate the proposed model.

1. Introduction

Identifying changes in a series of behaviors or events is crucial for a company to survive in a dynamic and competitive business environment. Especially because of advancements in data storage technology, numerous enterprises have constructed information systems, such as e-commerce systems and gamification-based apps, to serve customers and to simultaneously collect their behavioral data. The speed of data generation is high and existing technology can be used to collect abundant, varied, and instantaneous big data (i.e., data with high volume, velocity, and variety, the 3Vs) [5]. Managers can use these data to instantly achieve frequent and meaningful observations of customer behavior [4], for example, online consumer reviews [26].

Despite acquiring customer behavior data, discovered patterns, and their implied knowledge, managers remained concerned about one paramount question: "Is the current knowledge still valid or already outdated?" Marketers always release new and short-term promotions to promote products or services in today's competitive market. Thus, customers might rapidly change their preferences because of the influence from advertising or word of mouth from their friends; that is, changes in customer behaviors will always occur [18]. Therefore, managers need to quickly and often identify changes in customer behavior, renew their knowledge, and make timely and accurate responses to the change in order to adapt to the dynamic business environment.

One form of knowledge can be represented as follows: $< (a_1), \cdots, (a_k), \cdots, (a_m) > \rightarrow c$, where a_k represents an event (or a customer behavior) at the kth time point; m is the length of the temporal sequence; and c represents the resulting class. Taking a retail behavioral sequence as an example, $<$ (*watches*)(*jewelery*)(*antiques*) $> \rightarrow$ *VIP*, means that customers belonging to the VIP class have the following sequential behaviors: Purchasing watches occurs before purchasing jewelery, and finally, antique purchasing occurs. This type of sequence is called a *classifiable sequential pattern* (CSP) because the sequence is connected to a class with an arrow; it reveals the relationships between sequential causes (purchases of watches, jewelery, and antiques) and the effect (the given VIP membership). However, the previous year's pattern is substituted with $<$ (*smartphone*)(*jewelery*)(*antiques*) $> \rightarrow$ *VIP*. If a manager cannot recognize the change in a timely manner, he/she will reach an inaccurate conclusion with respect to the current trend: The VIP customers now purchase smart phones as their first priority. Without renewing this knowledge, managers might formulate inappropriate strategies to improve their services or designed promotions.

The primary aim in knowledge discovery and data mining is to use practical tools to discover behavioral patterns. Hence, numerous data-mining techniques have been proposed for producing useful behavioral knowledge in commerce, such as cross-selling in fuzzy time-interval sequential patterns [7], customer profiling [38], product

* Corresponding author.
E-mail address: ptyang@thu.edu.tw (P.-T. Yang).

https://doi.org/10.1016/j.dss.2017.11.007
0167-9236/

bundling [41], recency, frequency, and monetary (RFM) sequential patterns [8], and role change patterns in social networks [20]. Of the numerous data-mining techniques, CSP mining plays crucial role in assisting managers to identify implied sequential relationships between causes and effects [21].

To the authors' knowledge, no study has addressed the issue of change mining in CSPs. The contributions of this study can be summarized in three points:

1. This study proposes a novel change mining model, *SeqClassChange*, by inheriting the models of Refs. Song et al. [28], Tsai and Shieh [32], Huang [17] and Huang et al. [16], to identify the changes in CSPs at different time periods.
2. This study proposes a CSP similarity computation index (CSCI) to emphasize the cause-and-effect relationships between two CSPs.
3. *SeqClassChange* uses the CSP definition in Refs. Lesh et al. [21] and Tseng and Lee [36], rather than that of Zhao et al. [43], which uses itemsets instead of items at a single time point. That is, the proposed model allows multiple events (customer behaviors) to occur at each time-point in a CSP, which can help managers to model complex customer behavior.

Based on the above points, managers can identify more complex implications of cause-and-effect relationships and their related changes, and in order to make better decisions. Moreover, a real-world dataset is used to show the model's effectiveness and usefulness.

The remainder of this paper is organized as follows. Section 2 reviews related works including change mining and sequence-based classifications. Section 3 defines the *SeqClassChange* model for mining changes in CSPs. Section 4 provides the experimental results of the proposed model. Section 5 offers the conclusions.

2. Related work

Before the model for mining changes in CSPs is presented later in this paper, this section provides some basic research background. The information is separated into two different but correlated parts, namely change mining and sequence classifications, which are discussed in Sections 2.1 and 2.2, respectively. Finally, a fully integrated discussion is provided in Section 2.3.

2.1. Change mining

The core of the change mining problem is the association rule mining problem. The Apriori algorithm [1] was the original algorithm used to solve the association rule mining problem. It executes a number of iterations and the nth iteration can generate a set of n-length frequent patterns. The key to the Apriori algorithm is its generate-and-test (candidate generation and pruning) strategy. The new candidate itemsets in the current iteration are generated from the frequent itemsets in the previous iteration. Subsequently, the sequential version of the Apriori algorithm is also proposed in GSP [29].

Previous works have investigated change mining using two types of approaches: The maintenance approaches and the comparison approaches. The maintenance approaches adjust discovered patterns when new data, new transactions, or new customers are added to an original database, which improves the accuracy of the patterns in a dynamic environment [9,11,25]. This type of approach maintains existing knowledge but does not provide any extra information; that is, it cannot reveal the evolutionary process when customer behavior changes; therefore, a second type of approach has been proposed: the comparison approaches.

The comparison approaches recognize changes between different databases or within the same database at distinct time periods. The first study of change mining was conducted by Liu et al. [23]. The study stemmed from the post-analysis of learned rules [22], which devised an approach of change mining in the context of decision trees for six identifying behavioral changes of customers. Song et al. [28] developed an overall architecture including three phases that detects changes of customer behavior in association rules from databases. They investigated possible types of change on the basis of previous research, and summarized three types of change: (1) emerging patterns, (2) unexpected changes, and (3) added/perished rules. They then used similarity and difference measures to identify the changed rules and to rank them according to the degree of change. Chen et al. [6] integrated customer behavioral variables (recency, frequency, and monetary), demographic variables, and transaction databases to propose a method of mining changes in customer behavior. They also followed the definitions of the three types of change proposed in Song et al. [28] to identify change patterns. In addition, Bala [3] contended that the work by Chen et al. [6] did not consider products in the conditional part of a pattern depicted by association rules. Therefore, he proposed a change-mining approach, in which the conditional part may contain products or items. In terms of other types of technique in the comparison approaches, Au and Chan [2] introduced a fuzzy technique to identify association rules over time and generalized the problem so that different fuzzy data-mining techniques could be used to address the issue of change.

To identify changes in sequential patterns, Tsai and Shieh [32], using the architecture in Song et al. [28], first proposed a change detection framework to observe the dynamic alternation of sequential patterns between two time-periods and divided them into three significant types of change patterns, based on those in Song et al. [28]. Regarding discerning changes in fuzzy-based sequential patterns, *MineFuzzChange* [17] was the first change-mining model that discussed the change of customer behavior in fuzzy time-interval sequential patterns. Subsequently, a novel change mining model for detecting change in another type of sequential pattern, fuzzy quantitative sequential patterns [16], was proposed.

There is another issue which is similar to the idea of change mining, called *concept drift* [12,37][1]. However, the major concern of concept drift is to discuss the change of a predictive model in the conditional distribution of the output (i.e. target variable) given the input (input features). Specially, the distribution of the input may stay unchanged. Therefore, the past studies have proposed various adaptive learning algorithms for handling the issue. Since the traditional change mining does not concern on the distribution of the input or the output, this study will not address the issue of concept drift in advance.

2.2. Sequence classifications

Sequence classification, a sequence learning category [30], is used to determine whether a sequence is legitimate or is assigning class labels to new sequences. Previous studies have proposed different approaches to resolving the problems involved in sequence classifications, including decision trees, artificial neural networks, naive Bayes, k-nearest neighbors, the hidden Markov model, and support vector machines [15]. The investigation of sequence classification can be applied to numerous real-life circumstances such as protein function prediction [13], earthquake or typhoon prediction, text classification [39], debt detection in social security [43], and the class label prediction of new customers from temporal customer data [31].

The above approaches are algorithmic, statistical, or bionic methods. However, there is another branch of sequence classification: the pattern-based methods, which follow the generate-and-test (candidate generation and pruning) strategy [15], inspired by

[1] The authors thank the reviewer proffered the issue.