Knowledge-Based Systems xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-view attribute reduction model for traffic bottleneck analysis

X.D. Yue [a,*], L.B. Cao [b], D.Q. Miao [c], Y.F. Chen [c], B. Xu [d]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai 201804, China
[b] Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia
[c] School of Electronics and Information, Tongji University, Shanghai 200092, China
[d] School of Transportation Engineering, Tongji University, Shanghai 200092, China

ARTICLE INFO

Article history:
Received 15 October 2014
Received in revised form 19 March 2015
Accepted 20 March 2015
Available online xxxx

Keywords:
Pattern mining
Traffic bottleneck
User preference

ABSTRACT

In the field of traffic bottleneck analysis, it is expected to discover traffic congestion patterns from the reports of road conditions. However, data patterns mined by existing KDD algorithms may not coincide with the real application requirements. Different from academic researchers, traffic management officers do not pursue the most frequent patterns but always hold multiple views for mining task to facilitate traffic planning. They expect to study the correlation between traffic congestion and various kinds of road properties, especially the road properties easily to be improved. In this multi-view analysis, each view actually denotes a kind of user preference of road properties. Thus it is required to integrate user-defined attribute preferences into pattern mining process. To tackle this problem, we propose a multi-view attribute reduction model to discover the patterns of user interests. In this model, user views are expressed with attribute preferences and formally represented by attribute orders. Based on this, we implement a workflow of multi-view traffic bottleneck analysis, which consists of data preprocessing, preference representation and congestion pattern mining. We validate our approach based on the reports of road conditions from Shanghai. Experimental results show that the resultant multi-view mining outcomes are effective for analyzing congestion causes and traffic management.

1. Introduction

In the field of intelligent transportation, *knowledge discovery and data mining* (KDD) methods have been widely utilized to analyze various kinds of traffic data to construct decision support systems for traffic management [7,20,41,42]. For these data-driven traffic management systems, one of the most important tasks is analyzing the causes of traffic bottlenecks and taking action to alleviate congestion [2,17,24]. The data analysis of traffic bottlenecks are generally performed on either spatiotemporal data [15,16] or traffic reports of road conditions [3]. Specifically, for the traffic reports of road conditions, it is expected to analyze traffic bottlenecks through discovering congestion patterns from the table-formed data. To achieve this, most existing works directly apply the methodologies of association rule mining on traffic reports to obtain congestion patterns. Techniques of association rule mining and association analysis are employed to predict traffic network flows [11]. The algorithms of frequent pattern mining are used to discover simultaneously congested link-sets in a road network [28]. Additionally, a strategy of association rule acquisition based on group decision-making is proposed to identify the traffic states of regional road sections [14].

Depending on the above association rule/pattern mining methods, we can obtain abundant data patterns from traffic reports. However, lots of these patterns may be out of users' interests and not actionable enough to support traffic management. Because most existing data mining methods just focus on statistical significance of attributes, such as cooccurrence and discernibility to generate patterns but neglect user preferences and requirements. Without considering user preferences and requirements, the data mining algorithms provide same data patterns for different users and thus actually produce single-view analysis results. For real applications, the single-view analysis results are not sufficient and flexible enough to implement a solution for traffic improvement. Thus it is necessary to design user-oriented data mining methods to analyze traffic reports from multiple views.

Especially for the applications of traffic bottleneck analysis, traffic officers are eager to study the causes of traffic congestion from different perspectives. First, they hold the views to facilitate traffic planning. The discovered patterns should indicate the relationship between traffic congestion and the road properties that are easy to

* Corresponding author.
E-mail address: yswantfly@gmail.com (X.D. Yue).

http://dx.doi.org/10.1016/j.knosys.2015.03.022
0950-7051/© 2015 Elsevier B.V. All rights reserved.

be improved. Applying traditional data mining algorithms to road condition reports, the obtained congestion patterns generally consist of the attributes about road construction, such as road width, length and average delay. But in modern cities, these factors are always difficult to change, because reconstructing roads are really high-cost tasks. Thus officers expect to find the dependence between traffic congestion and the controllable factors, such as car/bike separators, zebra crossing, and bus lanes. Second, officers hold the views to involve experts' experiences into bottleneck analysis. The road properties being considered important for congestion formation should have high priority to emerge in the final patterns. The experts' experiences are helpful to discover, employ and interpret the actionable patterns.

As mentioned above, each view for traffic bottleneck analysis actually denotes a kind of preference of road properties. To achieve the multi-view analysis, it is required to integrate user-defined attribute preferences into pattern mining process. In other words, the desired data mining methods should be able to reduce and organize data attributes according to user preferences to generate the final patterns [8,18]. Aiming at the problems, we propose a multi-view attributes reduction model to discover the congestion patterns for traffic bottleneck analysis. The reduction model formulates user views with attribute preferences and extracts data patterns referring to the user-defined attribute priorities. Applying the multi-view reduction model to the traffic reports of road conditions, we can discover the congestion patterns of various kinds of user interests. The multi-view mining outcomes can induce comprehensive and practicable knowledge and lead to an overall analysis of traffic bottleneck causes. Our contributions are summarized as follows.

- *Propose a multi-view attribute reduction model to discover patterns of user interests.* In the model, user views are expressed with attribute preferences and formally represented by attribute orders. To implement the order-based attribute reduction, a data structure of 2D linked list is designed to storage item-discerning elements and discernibility thresholds are set to filter out trivial attributes in patterns.
- *Propose a workflow for traffic bottleneck analysis based on the multi-view reduction model.* The workflow consists of data collection and preprocessing of road conditions, user preference representation and attribute reduction for congestion pattern mining. It supports users to analyze the causes of urban traffic congestion from the views of traffic management.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 introduces a multi-view attribute reduction model, which includes user view representation, reduction algorithm implementation and theoretic model analysis. Based on the multi-view reduction model, Section 4 proposes a user-oriented workflow to analyze the causes of traffic bottlenecks. In Section 5, experimental results validate our approach is effective for overall traffic congestion analysis and customized knowledge discovery. The work conclusion is given in Section 6.

2. Related work

The basic idea of this work originates from the methodology of *actionable knowledge discovery* (AKD), which aims to find the actionable patterns which are friendly enough for business people to interpret, validate and action [5,33,35]. The key focus of actionable pattern mining is to involve user factors into data mining workflow [1,9]. Different from the traditional pattern evaluation of statistical significance, the evaluation of actionable patterns consists of the objective and subjective interestingness measures, which recognize to what extent a pattern is of interest to particular user preferences [19,23,31]. Specifically, the probability-based belief was used to describe user confidences of unexpected rules and the profit and utility mining frameworks were designed to measure the business values of patterns [6,34]. Based on the revised interestingness measures, domain-driven data mining strategies were further proposed to discover the workable knowledge for real applications [4,40]. In this paper, we expect to instantiate an actionable pattern mining model to cater for multiple user views. This model is constructed based on the process of attribute reduction.

As an inductive learning tool, *attribute reduction* extracts valuable patterns from table-formed information systems through reducing redundant attributes to attribute reducts [22,25,27]. Specifically, attribute reduction algorithms are good at processing uncertain data [21,30,36] and efficient to discover the rule-type knowledge from data tables [26,39]. The table-formed information systems are defined as $IS = (U, C, f, V)$, in which U is a finite set of data items, called the universe, C is a finite set of attributes to depict items, V denotes the domain of attribute values, and f is the mapping from U to V, which assigns particular attribute values to items. For classification tasks, we consider the information systems with decision attributes D, $DS = (U, C \cup D, f, V)$. Such systems are called *decision systems*, in which the attributes of C are viewed as conditions. In a decision system DS, for any $x, y \in U$, if $C(x) = C(y) \Rightarrow D(x) = D(y)$, system DS is *consistent*, otherwise it is *inconsistent*. In this paper, we just consider the consistent decision systems and assume D consists of a single attribute $D = \{d\}$, d is a decision attribute for labeling the class of each object. Example 1 illustrates a decision system about customer evaluation of cars.[1]

Example 1. Decision system 'Car evaluation': conditional attributes 'Buying price', 'Maintenance cost', 'Number of doors', 'Capacity of persons', 'Size of luggage boot', 'Safety of cars' depict the properties of cars and the decision attribute 'Accepted' reflects the evaluation of customers (Accepted or Unaccepted). The details are presented in Table 1.

In decision systems, the classification ability of a conditional attribute set A can be evaluated through constructing its *positive region* relative to the decision d. The positive region $POS_A(U/d)$ consists of all the objects that can be correctly classified with the attributes A. Attribute reduction aims to reduce redundant conditional attributes to an *attribute reduct* while preserve a certain classification property. Attribute reduct can be formally defined from the view of positive region preservation.

Definition 1. *Attribute Reduct.* Given a decision system $DS = (U, C \cup \{d\}, f, V)$, an attribute set $R \subseteq C$ is a d-reduct of C, iff

$$\begin{aligned} &(1) \quad POS_R(U/d) = POS_C(U/d) \\ &(2) \quad \forall r \in R,\ POS_{R-\{r\}}(U/d) \neq POS_R(U/d) \end{aligned} \tag{1}$$

Condition (1) requires a reduct R to have the same classification ability as the conditional attribute set C. For an attribute $r \in R$, if $POS_{R-\{r\}}(U/d) = POS_R(U/d)$, r is d-dispensable in R, otherwise r is d-indispensable. If all the attributes in R are d-indispensable, R is independent with respect to d, otherwise R is dependent. Obviously, condition (2) requires that a reduct R should be independent.

According to Definition 1, we notice that the reduct of a decision system may not be unique. The reduct of the minimum attributes is considered the *optimal reduct*. Such as in Example 1, attribute sets {Persons, Safety} and {Buying, Maintenance, Doors, Luggage

[1] The demo decision system is generated from the UCI dataset 'Car evaluation'.

Please cite this article in press as: X.D. Yue et al., Multi-view attribute reduction model for traffic bottleneck analysis, Knowl. Based Syst. (2015), http://dx.doi.org/10.1016/j.knosys.2015.03.022