Inter–Transactional Association Analysis for Prediction

Ling Feng
University of Twente, The Netherlands

Tharam Dillon
University of Technology Sydney, Australia

INTRODUCTION

The discovery of association rules from large amounts of structured or semi-structured data is an important data-mining problem (Agrawal et al., 1993; Agrawal & Srikant, 1994; Braga et al., 2002, 2003; Cong et al., 2002; Miyahara et al., 2001; Termier et al., 2002; Xiao et al., 2003). It has crucial applications in decision support and marketing strategy. The most prototypical application of association rules is market-basket analysis using transaction databases from supermarkets. These databases contain sales transaction records, each of which details items bought by a customer in the transaction. Mining association rules is the process of discovering knowledge such as, 80% of customers who bought diapers also bought beer, and 35% of customers bought both diapers and beer, which can be expressed as "*diaper* $\Rightarrow$ *beer*" (35%, 80%), where 80% is the confidence level of the rule, and 35% is the support level of the rule indicating how frequently the customers bought both diapers and beer. In general, an association rule takes the form $X \Rightarrow Y\,(s, c)$, where X and Y are sets of items, and s and c are support and confidence, respectively.

BACKGROUND

While the traditional association rules have demonstrated strong potential in areas such as improving marketing strategies for the retail industry (Dunham, 2003; Han & Kamer, 2001), their emphasis is on description rather than prediction. Such a limitation comes from the fact that traditional association rules only look at association relationships among items within the same transactions, whereas the notion of the transaction could be the items bought by the same customer, the atmospheric events that happened at the same time, and so on. To overcome this limitation, we extend the scope of mining association rules from such traditional intra-transactional associations to intertransactional associations for prediction (Feng et al., 1999, 2001; Lu et al., 2000). Compared to intratransactional associations, an intertransactional association describes the association relationships across different transactions, such as, if (company) A's stock goes up on day one, B's stock will go down on day two but go up on day four. In this case, whether we treat company or day as the unit of transaction, the associated items belong to different transactions.

MAIN TRUSTS

Extensions from Intratransaction to Intertransaction Associations

We extend a series of concepts and terminologies for intertransactional association analysis. Throughout the discussion, we assume that the following notation is used.

- A finite set of literals called items $I = \{i_1, i_2, \ldots, i_n\}$.
- A finite set of transaction records $T = \{t_1, t_2, \ldots, t_l\}$, where for $\forall t_i \in T$, $t_i \subseteq I$.
- A finite set of attributes called dimensional attributes $A = \{a_1, a_2, \ldots, a_m\}$, whose domains are finite subsets of nonnegative integers.

An Enhanced Transactional Database Model

In classical association analysis, records in a transactional database contain only items. Although transactions occur under certain contexts, such as time, place, customers, and so forth, such contextual information has been ignored in classical association rule mining, due to the fact that such rule mining was intratransactional in nature. However, when we talk about intertransactional associations across multiple transactions, the contexts of occurrence of transactions become important and must be taken into account.

Here, we enhance the traditional transactional database model by associating each transaction record with

Copyright © 2006, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

a number of attributes that describe the context within which the transaction happens. We call them *dimensional attributes*, because, together, these attributes constitute a multi-dimensional space, and each transaction can be mapped to a certain point in this space. Basically, dimensional attributes can be of any kind, as long as they are meaningful to applications. Time, distance, temperature, latitude, and so forth are typical dimensional attributes.

Multidimensional Contexts

An m-dimensional mining context can be defined through m dimensional attributes $a_1, a_2, \ldots, a_m$, each of which represents a dimension. When m=1, we have a single-dimensional mining context. Let $n_i = (n_i.a_1, n_i.a_2, \ldots, n_i.a_m)$ and $n_j = (n_j.a_1, n_j.a_2, \ldots, n_j.a_m)$ be two points in an m-dimensional space, whose values on the m dimensions are represented as $n_i.a_1, n_i.a_2, \ldots, n_i.a_m$ and $n_j.a_1, n_j.a_2, \ldots, n_j.a_m$, respectively. Two points n_i and n_j are equal, if and only if for $\forall k\,(1 \le k \le m)$, $n_i.a_k = n_j.a_k$. A relative distance between n_i and n_j is defined as $\Delta\langle n_i, n_j\rangle = (n_j.a_1 - n_i.a_1, n_j.a_2 - n_i.a_2, \ldots, n_j.a_m - n_i.a_m)$. We also use the notation $\Delta_{(d1, d2, \ldots, dm)}$, where $d_k = n_j.a_k - n_i.a_k\,(1 \le k \le m)$, to represent the relative distance between two points n_i and n_j in the m-dimensional space.

Besides, the absolute representation $(n_i.a_1, n_i.a_2, \ldots, n_i.a_m)$ for point n_i, we also can represent it by indicating its relative distance $\Delta\langle n_0, n_i\rangle$ from a certain reference point n_0, (i.e., $n_0 + \Delta\langle n_0, n_i\rangle$, where $n_i = n_0 + \Delta\langle n_0, n_i\rangle$). Note that n_i, $\Delta\langle n_0, n_i\rangle$, and $\Delta_{(ni.a1-n0.a1,\ ni.a2-n0.a2,\ \ldots,\ ni.am-n0.am)}$ can be used interchangeably, since each of them refers to the same point n_i in the space. Let $N = \{n_1, n_2, \ldots, n_u\}$ be a set of points in an m-dimensional space. We construct the smallest reference point of N, n_*, where for $\forall k\,(1 \le k \le m)$, $n_*.a_k = \min(n_1.a_k, n_2.a_k, \ldots, n_u.a_k)$.

Extended Items (Transactions)

The traditional concepts regarding item and transaction can be extended accordingly under an m-dimensional context. We call an item $i_k \in I$ happening at the point $\Delta_{(d1, d2, \ldots, dm)}$, (i.e., at the point $(n_0.a_1 + d_1, n_0.a_2 + d_2, \ldots, n_0.a_m + d_m)$), an extended item and denote it as $\Delta_{(d1, d2, \ldots, dm)}(i_k)$. In a similar fashion, we call a transaction $t_k \in T$ happening at the point $\Delta_{(d1, d2, \ldots, dm)}$ an extended transaction and denote it as $D_{(d1, d2, \ldots, dm)}(t_k)$. The set of all possible extended items, I_E, is defined as a set of $\Delta_{(d1, d2, \ldots, dm)}(i_k)$ for any $i_k \in I$ at all possible points $\Delta_{(d1, d2, \ldots, dm)}$ in the m-dimensional space. T_E is the set of all extended transactions, each of which contains a set of extended items, in the mining context.

Normalized Extended Item (Transaction) Sets

We call an extended itemset a *normalized extended itemset*, if all its extended items are positioned with respect to the smallest reference point of the set. In other words, the extended items in the set have the minimal relative distance 0 for each dimension. Formally, let $I_e = \{\Delta_{(d1,1, d1,2, \ldots, \sim d1,m)}(i_1), \Delta_{(d2,1, d2,2, \ldots, d2,m)}(i_2), \ldots, \Delta_{(dk,1, dk,2, \ldots, dk,m)}(i_k)\}$ be an extended itemset. I_e is a normalized extended itemset, if and only if for $\forall j\,(1 \le j \le k)\ \forall i\,(1 \le i \le m)$, $\min(d_{j,i}) = 0$.

The normalization concept can be applied to an extended transaction set as well. We call an extended transaction set a *normalized extended transaction set*, if all its extended transactions are positioned with respect to the smallest reference point of the set. Any non-normalized extended item (transaction) set can be transformed into a normalized one through a normalization process, where the intention is to reposition all the involved extended items (transactions) based on the smallest reference point of this set. We use I_{NE} and T_{NE} to denote the set of all possible normalized extended itemsets and normalized extended transaction sets, respectively. According to the above definitions, any superset of a normalized extended item (transaction) set is also a normalized extended item (transaction) set.

Multidimensional Intertransactional Association Rule Framework

With the above extensions, we are now in a position to formally define intertransactional association rules and related measurements.

Definition 1

A multidimensional intertransactional association rule is an implication of the form $X \Rightarrow Y$, where

(1) $X \subset I_{NE}$ and $Y \subset I_E$;
(2) The extended items in X and Y are positioned with respect to the same reference point;
(3) For $\forall \Delta_{(x1, x2, \ldots, xm)}(i_x) \in X$, $\forall \Delta_{(y1, y2, \ldots, ym)}(i_x) \in Y$, $x_j \le y_j$ $(1 \le j \le m)$;
(4) $X \cap Y = \varnothing$.

Different from classical intratransactional association rules, the intertransactional association rules capture the occurrence contexts of associated items. The first clause

4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/inter-transactional-association-analysis-prediction/10678

Related Content

Knowledge Discovery with Artificial Neural Networks
Juan R. Rabuñal Dopico, Daniel Rivero Cebrian, Julián Dorado de la Calleand Nieves Pedreira Souto (2005). *Encyclopedia of Data Warehousing and Mining (pp. 669-673).*
www.irma-international.org/chapter/knowledge-discovery-artificial-neural-networks/10681

Multi-Label Classification: An Overview
Grigorios Tsoumakasand Ioannis Katakis (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 64-74).*
www.irma-international.org/chapter/multi-label-classification/7632

Privacy-Preserving Data Mining on the Web: Foundations and Techniques
Stanley R.M. Oliveiraand Osmar R. Zaiane (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 50-63).*
www.irma-international.org/chapter/privacy-preserving-data-mining-web/7631

Humanitites Data Warehousing
Janet Delve (2005). *Encyclopedia of Data Warehousing and Mining (pp. 570-574).*
www.irma-international.org/chapter/humanitites-data-warehousing/10662

Why General Outlier Detection Techniques Do Not Suffice for Wireless Sensor Networks
Yang Zhang, Nirvana Meratniaand Paul Havinga (2010). *Intelligent Techniques for Warehousing and Mining Sensor Network Data (pp. 136-158).*
www.irma-international.org/chapter/general-outlier-detection-techniques-not/39544