Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Measuring influence of an item in a database over time

Jhimli Adhikari [a,*], P.R. Rao [b]

[a] Department of Computer Science, Narayan Zantye College, Bicholim, Goa 403 529, India
[b] Department of Computer Science and Technology, Goa University, Goa 403 206, India

ARTICLE INFO

*Article history:*
Received 15 July 2008
Received in revised form 6 October 2009
Available online 21 October 2009

Communicated by W. Pedrycz

*Keywords:*
Confidence
Overall association
Overall influence
Synthesis of influence

ABSTRACT

Influence of items on some other items might not be the same as the association between these sets of items. Many tasks of data analysis are based on expressing influence of items on other items. In this paper, we introduce the notion of an overall influence of a set of items on another set of items. We also propose an extension to the notion of overall association between two items in a database. Using the notion of overall influence, we have designed two algorithms for influence analysis involving specific items in a database. As the number of databases increases on a yearly basis, we have adopted incremental approach in these algorithms. Experimental results are reported for both synthetic and real-world databases.

© 2009 Elsevier B.V. All rights reserved.

## 1. Introduction

Every time a customer interacts with business, we have an opportunity to gain strategic knowledge. Transactional data contains a wealth of information about customers and their purchase patterns. In fact, these data could be one of the most valuable assets, when used wisely. This has been recognized a long time ago by many large organizations such as supermarkets, insurance companies, healthcare organizations, telecommunications, and banks. These organizations have spent significant resources for collecting and analyzing transactional data. Many applications are based on inherent knowledge present in a database (Gary and Petersen, 2000; Wu et al., 2005; Adhikari et al., 2009). Such applications could be dealt with mining databases (Han et al., 2000; Agrawal and Srikant, 1994; Savasere et al., 1995). As a database changes over time, the inherent knowledge also changes. Therefore in the competitive market, knowledge-based decisions are more appropriate. Data mining algorithms are effective tools to support making such decisions. Data mining algorithms often extracts different patterns from a database. Some examples of patterns in a database are frequent item sets (Agrawal et al., 1993), association rules (Agrawal et al., 1993), negative association rules (Wu et al., 2004), Boolean expressions induced by itemset (Adhikari and Rao, 2007b) and conditional patterns (Adhikari and Rao, 2008a). Nevertheless, there are some applications for which association-based analysis might be inappropriate. For example, an organization might deal with a large number of items with its customers. The company might be interested in knowing how the purchase of a particular item affects the purchase of some other item. In this paper, we study such influences based on transactional time-stamped database.

Many companies transact a large number of products (items) with their customers. It might be required to perform data analyses involving different items. Such analyses might originate from different applications. One such analysis is identifying stable items (Adhikari et al., 2009) in databases over time. It could be useful in devising strategies for a company. Little work has been reported on data analyses over time. In this paper, we present another application involving different items in a database over time.

Consider a company that collects a huge amount of transactional data on a yearly basis. Let $DT_i$ be the database corresponding to the $i$th year, $i = 1, 2, \ldots, k$. Each of these databases corresponds to a specific period of time. Thus, one could call these time databases. Each time database is mined using a traditional data mining technique (Adhikari and Rao, 2007a). In this application, we will deal with itemsets in a database. An itemset is a set of items in the database. Let $I$ be the set of all items in the time databases. Each itemset $X$ in a database $D$ is associated with a statistical measure, called *support* (Agrawal et al., 1993), denoted by $supp(X, D)$. The support of an itemset is defined as the fraction of transactions containing the itemset.

Solutions to many problems are based on the study of relationships among variables. We will see later that the study of influence of a set of variables on another set of variables might not be the same as the association between these two sets of variables. Association analysis among variables has been studied well (Agrawal et al., 1993; Adhikari and Rao, 2007a, 2008b,c; Brin et al., 1997; Shapiro, 1991). In the context of studying association among

[*] Corresponding author. Fax: +91 0832 2361377.
*E-mail addresses:* jhimli_adhikari@yahoo.co.in (J. Adhikari), pralhaad@rediffmail.com (P.R. Rao).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.10.010

variables using association rules one could conclude that the confidence of the association rule gives positive influence of antecedent on the consequent of the association rule. Such positive influences might not be sufficient for many data analyses.

Consider an established company possessing data over 50 consecutive years. Generally, the sales of a product vary from one season to another season. Also, a season re-appears on a yearly basis. Thus, we divide the entire database into a sequence of yearly databases. In this context, a yearly database could be considered as a time database. In this study, we estimate the influence of item $x$ on $y$, for $x$, $y \in I$, where $I$ is the set of all items in database $D$. In Section 3, we define the concept of influence of an itemset on another itemset.

An itemset could be viewed as a basic type of pattern in a database. Different types of pattern in a database could be derived from itemset patterns. For example, frequent itemset, association rule, negative association rule, Boolean expression induced by itemset and conditional pattern are examples of derived patterns in a database. Few applications have been reported on analysis of patterns over time. In this paper, we wish to study the influence of an item on a specific item/a set of specific items in a database.

Most of the association analyses are based on a positive association between variables. Such positive association gives rise to positive influence of variables on other variables. Most of the real databases are large and sparse. In such cases an association analysis using positive influence might not be appropriate, if the overall influence of former variable on latter variable becomes negative. Thus, the concept of overall influence needs to be introduced.

The rest of the paper is organized as follows: in Section 2, we extend the notion of overall association between two items in a database. In Section 3, we introduce the notion of overall influence of an itemset on another itemset in a database. We study various properties of proposed measures. Also, we introduce the notion of overall influence of an item on a set of specific items in a database. In addition, we discuss the motivation of the proposed problem in this section. We state our problem in Section 4. We discuss work related to proposed problem in Section 5. In Section 6, we design an algorithm to measure the overall influence of an item on another item (incrementally). In addition, we design another algorithm of overall influence of an item on a set of specific items (incrementally). Experimental results are provided in Section 7. We conclude the paper in Section 8.

## 2. Association between two itemsets

Adhikari and Rao (2007a) have proposed a measure denoted by *OA*, for computing an overall association between two items in a market basket data. Using positive association *PA* between two items (Adhikari and Rao, 2007a), one could extend positive association between two itemsets in a database as follows:

$$PA(X, Y, D) = \frac{\#\text{transaction containing both } X \text{ and } Y, D}{\#\text{transaction containing at least one of } X \text{ and } Y, D},$$

where $X$ and $Y$ are itemsets in database $D$ and "#$P$, $D$" is the number of transactions in $D$ that satisfy the predicate $P$.

Similarly, negative association *NA* between two items (Adhikari and Rao, 2007a) could be extended as follows:

$$NA(X, Y, D) = \frac{\#\text{transaction containing exactly one of } X \text{ and } Y, D}{\#\text{transaction containing at least one of } X \text{ and } Y, D},$$

where $X$ and $Y$ are itemsets in database $D$.

Using *PA* and *NA*, *OA* between two itemsets $X$ and $Y$ in database $D$ could be defined as follows:

$$OA(X, Y, D) = PA(X, Y, D) - NA(X, Y, D). \tag{1}$$

If $OA(X, Y, D)$ is positive, negative or zero then all the items in $X$ together and all the items in $Y$ together are positively, negatively or independently associated in $D$, respectively. We illustrate different types of association in the following example.

**Example 1.** Let database $D_1$ contain the following transactions: $\{a, d, e\}, \{a, b, c, d, g\}, \{a, b, e, g\}, \{b, c, g\}, \{d, e, g\}, \{b, e, f\}, \{c, d, e, f\}$, $\{a, b, c, d, f, g\}$, and $\{a, b, c, d, e\}$. We find here overall association between itemsets $X$, and $Y$, for some $X$, $Y$ in $D_1$. In Table 1, supports of some itemsets are given below.

Here $PA(\{a, b\}, \{c, d\}, D_1) = 3/5$ and $NA(\{a, b\}, \{c, d\}, D_1) = 2/5$. Therefore, $OA(\{a, b\}, \{c, d\}, D_1) = 1/5$. In Table 2, overall associations are given.

In Table 2, we observe that the *OA* value between $\{a, b\}$ and $\{c, d\}$ as well as $\{a, c\}$ and $\{b, d\}$ are positive. But, the *OA* value between $\{c\}$ and $\{d, e\}$ is negative.

## 3. Concept of influence

Let $X$ and $Y$ be two itemsets in database $D$. We wish to find influence of $X$ on $Y$ in $D$. In the above section, we have proposed overall association between two itemsets. The influence of $X$ on $Y$ seems to be different from overall association between $X$ and $Y$.

Let $X = \{x_1, x_2, \ldots, x_p\}$ and $Y = \{y_1, y_2, \ldots, y_q\}$ be two itemsets in database $D$. The influence of $X$ on $Y$ could be judged by the following events: (i) whether a customer purchases all the items of $Y$ when they purchase all the items of $X$ and (ii) whether a customer purchases all the items of $Y$ when they do not purchase all the items of $X$. Such behaviors could be modeled using supports of $X \cap Y$ and $\neg X \cap Y$. The expression $supp(X \cap Y, D)/supp(X, D)$ measures the strength of positive association of $X$ on $Y$. The expression $supp(\neg X \cap Y, D)/supp(\neg X, D)$ measures the strength of negative association of $X$ on $Y$. Thus, the expressions $supp(X \cap Y, D)/supp(X, D)$ and $supp(\neg X \cap Y, D)/supp(\neg X, D)$ could be important in measuring overall influence of $X$ on $Y$.

### 3.1. Influence of an itemset on another itemset

Let $X$ and $Y$ be the two itemsets in database $D$. The interestingness of an association rule $r_1$: $X \rightarrow Y$ could be expressed by its support and confidence (*conf*) measures (Agrawal et al., 1993). These measures are defined as follows. $supp(r_1, D) = supp(X \cap Y, D)$, and $conf(r_1, D) = supp(X \cap Y, D/supp(X, D)$. The measure $conf(r_1, D)$ could be interpreted as the fraction of transactions containing itemset $Y$ among the transactions containing $X$ in $D$. In other words, $conf(r_1, D)$ could be viewed as the *positive influence* (*PI*) of $X$ on $Y$. Let us consider the negative association rule $r_2$: $\neg X \rightarrow Y$. Confidence of $r_2$ in $D$ could be viewed as fractions of transactions containing $Y$ among the transactions containing $\neg X$. In other words, confidence of $r_2$ in $D$ could be viewed as *negative influence* (*NI*) of $X$ on $Y$. In similar to overall association defined in (1), one could define *overall influence* (*OI*) of $X$ on $Y$ in a database as follows:

**Table 1**
Supports of itemsets in $D_1$.

| Itemset($\{X\}$) | $\{a, b\}$ | $\{c, d\}$ | $\{a, c\}$ | $\{b, d\}$ | $\{d, e\}$ | $\{e, g\}$ |
|---|---|---|---|---|---|---|
| $supp(\{X\}, D_1)$ | 4/9 | 4/9 | 3/9 | 3/9 | 4/9 | 2/9 |

**Table 2**
Overall association between two itemsets in $D_1$.

| Itemset($\{X, Y\}$) | $\{\{a, b\}, \{c, d\}\}$ | $\{\{a, c\}, \{b, d\}\}$ | $\{\{c\}, \{d, e\}\}$ |
|---|---|---|---|
| $OA(X, Y, D_1)$ | 1/5 | 1 | −3/7 |

Download English Version:

https://daneshyari.com/en/article/536108

Download Persian Version:

https://daneshyari.com/article/536108

Daneshyari.com