Association Bundle Identification

Wenxue Huang
Generation5 Mathematical Technologies, Inc., Canada

Milorad Krneta
Generation5 Mathematical Technologies, Inc., Canada

Limin Lin
Generation5 Mathematical Technologies, Inc., Canada
Mathematics and Statistics Department, York University, Toronto, Canada

Jianhong Wu
Mathematics and Statistics Department, York University, Toronto, Canada

INTRODUCTION

An association pattern describes how a group of items (for example, retail products) are statistically associated together, and a meaningful association pattern identifies 'interesting' knowledge from data. A well-established association pattern is the *association rule* (Agrawal, Imielinski & Swami, 1993), which describes how *two sets of items* are associated with each other. For example, an association rule *A-->B* tells that 'if customers buy the set of product *A*, they would also buy the set of product *B* with probability greater than or equal to *c*'.

Association rules have been widely accepted for their simplicity and comprehensibility in problem statement, and subsequent modifications have also been made in order to produce more interesting knowledge, see (Brin, Motani, Ullman and Tsur, 1997; Aggarwal and Yu, 1998; Liu, Hsu and Ma, 1999; Bruzzese and Davino, 2001; Barber and Hamilton, 2003; Scheffer, 2005; Li, 2006). A relevant concept is the *rule interest* and excellent discussion can be found in (Shapiro 1991; Tan, Kumar and Srivastava, 2004). Huang et al. recently developed *association bundles* as a new pattern for association analysis (Huang, Krneta, Lin and Wu, 2006). Rather than replacing the association rule, the association bundle provides a distinctive pattern that can present meaningful knowledge not explored by association rules or any of its modifications.

BACKGROUND

Association bundles are important to the field of Association Discovery. The following comparison between association bundles and association rules support this argument. This comparison is made with focus on the association structure.

An *association structure* describes the structural features of an association pattern. It tells how many association relationships are presented by the pattern, and whether these relationships are asymmetric or symmetric, between-set or between-item. For example, an association rule contains one association relationship, and this relationship exists between two sets of item, and it is asymmetric from the rule antecedent to the rule consequent. However, the asymmetric between-set association structure limits the application of association rules in two ways. Firstly, when reasoning based on an association rule, the items in the rule antecedent (or consequent) must be treated as whole -a combined item, not as individual items. One can not reason based on an association rule that a certain individual antecedent item, as one of the many items in rule antecedent, is associated with any or all of the consequent items. Secondly, one must be careful that this association between the rule antecedent and the rule consequent is asymmetric. If the occurrence of the entire set of antecedent items is not deterministically given, for example, the only given information is that a customer has chosen the consequent items, not the antecedent items, it is highly probably that she/he does not chose any of the antecedent items. Therefore, for applications where between-item symmetric associations are required, for example, cross selling a group of items by discounting on one of them, association rules cannot be applied.

The association bundle is developed to resolve the above problems by considering the symmetric pair-wise between-item association structure. There are multiple association relationships existing in an association bundle - every two bundle-elements are associated with each other, and this between-element association is symmetric – there is no difference between the two associated items in terms of antecedence or consequence. With the symmetric between-element association structure, association bundles can be applied to applications where the asymmetric between-set association rules fail. Association bundles support marketing efforts where the sales improvement is expected on *every element* in a product group. One such example is the shelf management. An association bundle suggests that whenever and whichever an item *i* in the bundle is chosen by customers, every other item *j* in the bundle should possibly be chosen as well, thus items from the same bundle should be put together in the same shelf. Another example is the cross-selling by discounting. Every weekend retailers print on their flyers the discount list, and if two items have strong positive correlation, they should perhaps not be discounted simultaneously. With this reasoning, an association bundle can be used to do list checking, such that only one item in an association bundle will be discounted.

PRINCIPAL IDEAS

Let *S* be a transaction data set of *N* records, and *I* the set of items defining *S*. The *probability of an item k* is defined as *Pr(k) = |S(k)| / N,* where *|S(k)|* is the number of records containing the item *k*. The *joint probability of two items j and k* is defined as *Pr(j,k) = |S(j,k)| / N,* where *|S(j,k)|* is the number of records containing both *j* and *k*. The *conditional probability of the item j with respect to the item k* is defined as *Pr(j|k) = Pr(j) / Pr(j,k)*, and the *lift the item j and k* is defined as *Lift(j,k) = Pr(j,k) / (Pr(j)*Pr(k)).*

Definition. An association bundle is a group of items $b=\{i_1,\ldots,i_m\}$, a subset of *I*, that any two elements i_j and i_k of *b* are associated by satisfying that

(i). the lift for i_j and i_k is greater than or equal to a given threshold *L*, that is,

$$Pr(i_j, i_k) / (Pr(i_j) * Pr(i_k)) >= L;$$

(ii). both conditional probabilities between i_j and i_k are greater than or equal to a given threshold *T*, that is,

$$\mathrm{Pr}(i_j \mid i_k) >= \mathrm{T} \quad \text{and} \quad \mathrm{Pr}(i_k \mid i_j) >= \mathrm{T}.$$

An example is shown in the Figure 1 on association bundles. Figure 1 contains six tables. The first table shows the transaction data set, which is the one that used by Agrawal etc. (Agrawal, et. al., 1994) to illustrate the identification of association rules. The second and the third tables display the between-item conditional probability and lift values, respectively. The forth table displays the item pairs that have the conditional probability and lift values greater than or equal to the given thresholds, these item pairs are associated item pairs by definition. The fifth table shows the identified association bundles. For comparison, we display the association rules in the sixth table. A comparison between the association bundles and the association rules reveals that the item set {2,3,5} is identified as an association rule but not an association bundle. Check the fourth table we can see the item pair {2,3} and the item pair {3,5} actually have the lift values smaller than 1, which implies that they are having negative association with each other.

We further introduce association bundles in the following four aspects—association measure, threshold setting of measure, supporting algorithm, and main properties—via comparisons between association bundles and association rules.

Association Measure

The conditional probability (confidence) is used as the association measure for association rules (Agrawal, Imielinski & Swami, 1993), and later other measures are introduced (Liu, Hsu and Ma, 1999, Omiecinski 2003). Detailed discussions about association measures can be found in (Tan, Kumar and Srivastava, 2004). Association bundles use the between-item lift and the between-item conditional probabilities as the association measures (Huang, Krneta, Lin and Wu, 2006). The between-item lift guarantees that there is strong positive correlation between items: the between-item conditional probabilities ensure that the prediction