Association rules

The text and the exercises are part of the DBTech Virtual Workshop on BI. For more information, see http://www.dbtechnet.org

What are association rules?

Association rules express regularities that that exist in a dataset. Because a vast amount of different association rules can be derived from even a tiny dataset, interest is restricted to those that occur often and that predict with a high confidence. Let us now introduce some terminology:

An association rule is of the form A -> B, where A is called the **left hand side** of the rule and B is called the **right hand side** of the rule. The set of association rules is derived from a set of **transactions** in a database. Let us illustrate this with the dataset given in table 1. It contains two transactions, namely T1 and T2. Each transaction happens to contain three **item**s. For example, the items of transaction T1 are A, B and C.

T1	A	B	C
T2	B	C	D

Table 1: An example dataset containing transactions T1 and T2.

From the transaction T1 alone, it is possible to generate the set of association rules that is listed in Table 2:

A->B s:1 i.e. 50%, c:1 i.e. 100%	A->C s:1 i.e. 50%, c:1 i.e. 100%	A-> B, C s:1 i.e. 50%, c:1 i.e. 100%	B,C -> A s:1 i.e. 50%, c:1 i.e. 100%
B-> A s:1 i.e. 50%, c:1 i.e. 100%	B->C s:2 i.e. 100%, c:1 i.e. 100%	B-> C,A s:1 i.e. 50%, c:1/2 i.e. 50%	C,A->B s:1 i.e. 50%, c:1 i.e. 100%
C-> A s:1 i.e. 50%, c:1/2 i.e. 50%	C-> B s:2 i.e. 100% c:1 i.e. 100%	C-> A,B s:1 i.e. 50%, c:1/2 i.e. 50%	A,B ->C s:1 i.e. 50%, c:1 i.e. 100%

Table 2: The association that can be derived from the transaction T1 in Table 1. In addition to the rules, also their support (s) and confidence (c) values have been calculated. The support and confidence of the rules have been calculated based on the entire dataset consisting of the transactions T1 and T2. The values are expressed both as absolute values and as percentages.

Because the number of possible association rules that may be generated from a dataset is huge, it is important to restrict the interest to those that occur often and that predict with a high confidence. These two properties that describe the **interestingness of a rule** are called **support (s)** and **confidence (c)**. The support of a rule is simply the number of times it appears in a dataset. The confidence of a rule is the number of time that the left hand side and the right hand side of the rule appear together divided by the number of times that the left hand side of the rule appears in the dataset. More formally:

s(A->B) = frequency(A,B), where frequency(A,B) is the number of transactions that contain both A and B.

c(A->B) = frequency(A,B) / frequency(A).

Examples of support and confidence values of association rules are given in Table 2. The values have been calculated based on the dataset in Table 1.

How to derive association rules from a dataset?

As the amount of possible association rules is often huge and as association rule mining is usually performed on very large databases, special attention has been paid to developing efficient algorithms for association rule mining. As we saw in the example given in Table 2, it is possible to generate all possible association rules based on a dataset and then calculate the support and confidence values for each of them. However, this approach is very ineffective. A more effective approach is obtained by concentrating on the interestingness of the association rules right from the beginning. In the following, we will explain how this can be done.

In order to concentrate on the interestingness of the association rules right from the beginning and in order to avoid the creation of uninteresting association rules, we will for the time being ignore the distinction into the right hand side and the left hand side of a rule. Instead, we will just search for combinations of items that have a prespecified minimum support. The combinations of items are called **item sets**, and the combinations of items with a prespecified minimum support are called **frequent item sets**.

The method functions so that we first create the item sets that contain one item and that have the minimum support in the dataset. The next step is to create the item sets that contain two items based on the item sets created in the previous phase. Again, also the item sets of two have to have the minimum support. Bigger and bigger frequent item sets are created until all the possibilities given in the dataset are used. Let us illustrate this procedure with the dataset given in Table 1. Let us determine that the minimum support has to be 2, i.e. 100%.

Number of items	Item sets and their support
1	{B} s:2, {C} s:2
2	{B,C} s:2

Table 3. Item sets with a minimum support of 2 created from the dataset given in Table 1. Sets are denoted by curly brackets, {}.

The item sets with a minimum support value of 2 that can be created from the dataset of Table 1 are given in the Table 3 above. As we can see, no item sets of three items with a minimum support of 2 can be created. Now that we have the frequent item sets, we will proceed to develop the association rules based on these sets. The next step is to produce all possible association rules based on the frequent item sets, calculate the confidence values of the rules and keep those rules whose confidence value is high enough. Let us illustrate this with the example frequent item sets given in Table 3 and with the data given in Table 1. Let the minimum confidence value be 75%.

Association rule	support	confidence
B->C	2	1

C->B	2	1

Table 4. The association rules generated from the frequent item sets given in Table 3. The support and confidence of each rule is also shown.

As we can see from the example illustrated in Table 4, only frequent item sets with at least 2 items can be used to create association rules. If the minimum support is set to 1 (i.e. 100%) and the minimum confidence value to 75%, the rules that are created are the following: B->C and C->B.

Determining the association rules manually is not feasible when one has to deal with a real-life dataset. This is why several **algorithms** for the efficient generation of association rules have been created. The algorithms consist of two stages: generating the frequent item sets and from each frequent item set, determining the rules that that have the specified minimum confidence. One software that has several implementations of association rule mining algorithms is WEKA.

The above text is based on the pages 112 - 117 in Witten, Ian: Practical tools for Data Mining and the following articles from Wikipedia: “Association rules” and “Apriori algorithm”.

Exercises on Association Rule Mining

The exercises are part of the DBTech Virtual Workshop on BI.

Exercise 1. Basic association rule creation manually.

The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e. accuracy) 80%

Trans_id	Itemlist
T1	{K, A, D, B}
T2	{D, A C, E, B}
T3	{C, A, B, E}
T4	{B, A, D}