DEFINING INTERESTINGNESS FOR ASSOCIATION RULES

T. Brijs, K. Vanhoof, G. Wets

Abstract: Interestingness in Association Rules has been a major topic of research in the past decade. The reason is that the strength of association rules, i.e. its ability to discover ALL patterns given some thresholds on support and confidence, is also its weakness. Indeed, a typical association rules analysis on real data often results in hundreds or thousands of patterns creating a data mining problem of the second order. In other words, it is not straightforward to determine which of those rules are interesting for the end-user. This paper provides an overview of some existing measures of interestingness and we will comment on their properties. In general, interestingness measures can be divided into objective and subjective measures. Objective measures tend to express interestingness by means of statistical or mathematical criteria, whereas subjective measures of interestingness aim at capturing more practical criteria that should be taken into account, such as unexpectedness or actionability of rules. This paper only focusses on objective measures of interestingness.
Keywords: IJ ITA, formatting rules.

Introduction

The problem of finding association rules $X \Rightarrow Y$ was first introduced in 1993 by Agrawal, Imielinski and Swami [1993] as the data mining task of finding frequently co-occurring items in a large Boolean transactional database *D*. Typical applications include retail market basket analysis [Brijs et al., 1999; Brijs et al. 2000], item recommendation systems, cross-selling, loss-leader analysis, etc. In the classical framework, an association rule is considered to be interesting if its support (*s*) and confidence (*c*) exceed some user-defined minimum thresholds. *Support* is defined as the percentage of transactions in the data that contain all items in both the antecedent and the consequent of the rule, i.e. $P(X \cap Y) = \{X \cap Y\}/\{D\}$. *Confidence* on the other hand is an estimate of the conditional probability of *Y* given *X*, i.e. $P(X \cap Y)/P(X)$.

Several authors [Aggarwal & Yu, 1998], however, criticized the use of support and confidence for defining interesting associations. There are several reasons for this. First of all, it is not trivial to set good values for the minimum support and confidence thresholds. Optimally, given unlimited computing resources, these values should be dependent on the size of the data, the sparseness of the data and the particular problem under study. With respect to the size of the data, both the number of rows and columns in the data have an impact on the computing time and the number of association rules being generated. Indeed, for most association rule algorithms, computing time is known to be linear with the number of records in the database [Agrawal & Srikant, 1994]. Furthermore, given a particular percentage threshold for support, the absolute support for a rule will be totally different for a small or a large database. This also has an important impact on the statistical robustness of an association rule and is better known as sampling variability [DuMouchel & Pregibon, 2001; DuMouchel, 1999]. The idea is that association rules with low absolute support should be handled with care since small changes in the absolute support of an association rule with low support have a much greater impact than for an association rule with high absolute support. For example, for a rule with absolute support 2, an absolute support increase of 2 implies the rule to become twice as important as the original rule. In contrast, for a rule with absolute support of 2000, an absolute increase of 2 implies almost no change in the importance of the rule. With respect to the number of columns/items in the data, it is known that this may have a dramatic impact on the computing time, especially if the data is not sparse since the number of potential frequent candidates (and thus also computing time) will increase dramatically with the number of columns in the data. However, as long as the data are sparse, an increase in the number of columns in the database will not significantly increase the computing time due to the clever downward closure principle of frequent itemset mining [Agrawal & Srikant, 1994].

Furthermore, there is a fundamental critique in so far that the same support threshold is being used for rules containing a different number of items. Indeed, intuitively it is not clear why the same support threshold should apply for itemsets of size 2 or of size 7. Clearly, we expect the latter to occur much less frequently so, in some sense, it seems intuitive to specify different (i.e. lower) thresholds for itemsets of increasing size.

Finally, the nature of the problem under study may dictate the support and/or confidence thresholds that should be used. For instance, setting the support threshold too low may lead to rules for which the target group of customers of a particular marketing campaign based on those rules is too small. On the other hand, setting the threshold too high may lead to rules that are trivial for the retailer. Unfortunately, setting the right

values for minimum support and confidence today remains to be an unsolved problem in association rule mining.

Another limitation with regard to the support-confidence framework is that high confidence should not be confused with high correlation, neither with causality between the antecedent and the consequent of the rule. The former can be illustrated by the following example.

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Clearly, from this table it can be observed that *X* and *Y* are positively correlated and that *X* and *Z* are negatively correlated. Yet, if we calculate the support and confidence of the rules $X \Rightarrow Y$ (25%, 50%) and $X \Rightarrow Z$ (37.5%, 75%) it turns out that the second rule dominates the first both in terms of support and confidence. This example demonstrates that one should be very careful in defining the interestingness of rules in terms of support and/or confidence.

A second example illustrates why confidence can be misleading to define interesting rules. Suppose the following situation. Among 5000 customers:

- 3000 buy cola
- 3750 buy cheese
- 2000 both purchase cola and cheese

Then, the rule buy cola $\Rightarrow$ buy cheese (40%, 66%) could indicate a promising rule. However, although the rule has promising confidence, it is totally misleading since the baseline frequency of customers buying cheese is 75%. In other words, among all customers buying cola, the proportion of customers buying cheese is even lower than in the total group of customers. This example illustrates that one should always take into account the baseline frequency of the consequent of the rule when evaluating the interestingness of association rules.

Interestingness Measures

In the next paragraphs, we will provide an overview of most of the well-known objective interestingness measures, together with their advantages or disadvantages. Furthermore, all measures are symmetric measures, so the direction of the rule ($X \Rightarrow Y$ or $Y \Rightarrow X$) is not taken into account. The reason why we do not discuss a-symmetric measures is that, to our opinion, in retail market basket analysis it does not make sense to account for the direction of a rule since the concept of direction in association rules is meaningless in the context of causality. The interested reader is referred to Tan et al. [2001] for an overview of interestingness measures (both symmetric and a-symmetric) and their properties.

Lift / Interest

A few years after the introduction of association rules, researchers [Aggarwal & Yu, 1998; Brin et al., 1998] started to realize the disadvantages of the confidence measure by not taking into account the baseline frequency of the consequent. Therefore, the lift (also called interest) measure was introduced:

$$I = \frac{P(X \cap Y)}{P(X)P(Y)}$$

Since *P(Y)* appears in the denominator of the interest measure, the interest can be seen as the confidence divided by the baseline frequency of *Y*. The interest measure is defined over $[0, \infty[$ and its interpretation is as follows:

- If $I < 1$, then *X* and *Y* appear less frequently together in the data than expected under the assumption of conditional independence. *X* and *Y* are said to be negatively interdependent.
- If $I = 1$, then *X* and *Y* appear as frequently together as expected under the assumption of conditional independence. *X* and *Y* are said to be independent of each other.
- If $I > 1$, then *X* and *Y* appear more frequently together in the data than expected under the assumption of conditional independence. *X* and *Y* are said to be positively interdependent.

For instance, the interest value for the cola/cheese example equals I = 0.4/(0.6*0.75)=0.888 which is clearly below 1 and indicates that buying cola and buying cheese is negatively interdependent, as we expected.

There are, however, two important limitations to the interest measure [DuMouchel & Pregibon, 2001; DuMouchel, 1999]. The first one is related to the problem of sampling variability (see section Empirical Bayes Estimate). This means that for low absolute support values, the value of the interest measure may fluctuate heavily for small changes in the value of the absolute support of a rule. This problem is solved by introducing an Empirical Bayes estimate of the interest measure. The second problem is that the interest measure should not be used to compare the interestingness of itemsets of different size. Indeed, the interest tends to be higher for large itemsets than for small itemsets. The reason is that due to the conditional independence assumption in the denominator of the interest measure, the value in the denominator decreases much more rapidly than the value of the nominator when the number of items in the itemset increase. Therefore, the value of the interest will usually overestimate the interestingness of large itemsets.

Chi-square Test for Independency

A natural way to express the dependence between the antecedent and the consequent of an association rule $X \Rightarrow Y$ is the correlation measure based on the Chi-square test for independence [Brin et al., 1998]. Using the cola/cheese example again, the following contingency table can be derived from it:

	Buy cheese	Do not buy cheese	Total
Buy cola	2000	1000	3000
Do not buy cola	1750	150	2000
Total	3750	1150	5000

The chi-square test for independence is calculated as follows, with O_{xy} the observed frequency in the contingency table and E_{xy} the expected frequency (by multiplying the row and column total divided by the grand total):

$$\chi^2 = \sum_x \sum_y \frac{\left(O_{xy} - E_{xy}\right)^2}{E_{xy}} =$$

$$\frac{\left(2000 - \frac{3000*3750}{5000}\right)^2}{\frac{3000*3750}{5000}} + \frac{\left(1000 - \frac{3000*1150}{5000}\right)^2}{\frac{3000*1150}{5000}} + \frac{\left(1750 - \frac{2000*3750}{5000}\right)^2}{\frac{2000*3750}{5000}} + \frac{\left(150 - \frac{2000*1150}{5000}\right)^2}{\frac{2000*1150}{5000}} =$$

$$417.63 >> 3.84$$

For the p-value of 0.05 with one degree of freedom, the cutoff value equals 3.84. Consequently, buying cola and buying cheese can be considered as highly interdependent at the 95% confidence level. The correlation measure therefore seems an attractive alternative to the interest measure. However, the correlation measure has some important limitations with respect to using large data sets [Silverstein et al., 1998].

First of all, the Chi-square test rests on the normal approximation to the Binomial distribution. This approximation

breaks down when the expected values (E_{xy}) are small. More specifically, the Chi-square test should only be used when all cells in the contingency table have expected values greater than 1 and at least 80% of the cells have expected values greater than 5. In market basket data, however, these requirements are easily violated. Secondly, the values in the cell of the contingency table will typically be very unbalanced in the case of association rules. The reason is that the combination of non-existence of the items in the antecedent and consequent is usually much larger than the co-occurrence of its items. In other words, in real applications the upper left cell will be several orders of magnitude smaller than the lower right cell of the contingency table. This situation will usually invalidate the use of the Chi-square test for independence. Finally, the Chi-square test will produce larger values when the data set grows to infinity. Therefore, more items will tend to become significantly interdependent if the size of the dataset increases. The reason is that the Chi-square value depends on the total number of transactions, whereas the critical cutoff value only depends on the degrees of

freedom (which is equal to 1 for binary variables) and the desired significance level. Therefore, whilst comparison of Chi-squared values within the same data set may be meaningful, it is certainly not advisable to compare Chi-squared values across different data sets. In any of these cases, an exact test (like Fisher's exact test) to measure the significance of the interdependency is preferred over the Chi-square approximation.

The advantage of the chi-square measure, on the other hand, is that it takes into account all the available information in the data about the occurrence or non-occurrence of combinations of items, whereas the lift/interest measure only measures the co-occurrence of two itemsets, corresponding to the upper left cell in the contingency table.

Correlation Coefficient

The correlation coefficient (also known as the Φ-coefficient) measures the degree of linear interdependency between a pair of random variables. It is defined by the covariance between the two variables divided by their standard deviations:

$$\rho_{XY} = \frac{P(X \cap Y) - P(X)P(Y)}{\sqrt{P(X)(1-P(X))}\sqrt{P(Y)(1-P(Y))}}$$

where $\rho_{XY} = 0$ when X and Y are independent and ranges from [-1, +1].

Log-linear Analysis

A natural extension of the Chi-square test of independence between two-way contingency tables is the log-linear analysis [Agresti, 1996]. Log-linear analysis is suited to measure the interdependency between multi-way contigency tables. This kind of test is suited when we are not interested in finding the interdependency between the antecedent and the consequent of an association rule, but we are interested in the interdependency of individual items within an itemset. The log-linear model is one of the specialized cases of generalized linear models for Poisson-distributed data. Log-linear models are commonly used to evaluate multi-way contingency tables that involve three or more variables.

The basic strategy in log-linear modelling involves fitting models to the observed frequencies in the cross-tabulation of categorical variables. The models can then be represented by a set of expected frequencies. Models will vary in terms of the marginals they fit, and can be described in terms of the constraints they impose on the associations or interactions that are present in the data. Once expected frequencies are obtained, different models can be compared that are hierarchical to one another. The purpose is then to choose a preferred model, which is the most parsimonious model that fits the data. The choice of a preferred model is typically based on a formal comparison of goodness-of-fit statistics (likelihood ratio test) associated with models that are related hierarchically (i.e. models containing higher order terms also implicitly include all lower order terms). For instance, the fully-saturated log-linear model for two variables X and Y is:

$$\ln(F_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

where $\ln(F_{ij})$ is the log of the expected cell frequency of the cases for cell ij in the contingency table, μ is the overall mean of the natural log of the expected frequencies, λ represents 'effects' which the variables have on the cell frequencies, X and Y are the variables, and finally i and j refer to the categories within the variables. The above model is the fully-saturated model since it includes all possible one-way and two-way effects. In order to find a more parsimonious model that will isolate the effects best demonstrating the data patterns, a *non-saturated* model must be discovered. This can best be achieved by setting some of the effect parameters equal to zero. For instance, if the effects parameter λ_{ij}^{XY} is set to zero (i.e. we assume that X has no effect on Y and vice versa), the following unsaturated model is obtained:

$$\ln(F_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$$

Moreover, the unsaturated and the saturated model are hierarchically related, i.e. they are said to be *nested*. This is a very attractive feature since it validates the use of the likelihood ratio test (LRT). In fact, if F_{ij} represents the observed frequency, and f_{ij} the fitted frequency, then the likelihood ratio test is defined as:

$$LRT = 2\sum_i \sum_j F_{ij} \log\left(\frac{F_{ij}}{f_{ij}}\right)$$

The LRT test is distributed Chi-square with degrees of freedom equal to the number of cells minus the number of non-redundant parameters in the model. In other words, the degrees of freedom equal the number of λ parameters set equal to zero. In fact, the LRT tests the residual frequency not accounted for by the effects in the model. Thus, larger values for LRT indicate that the model does not fit the data well, and thus the model should be rejected. At this point, the LRT can be used to compare the saturated model with any other nested model:

$$LRT_{\text{difference}} = LRT_{\text{nested}} - LRT_{\text{saturated}}$$

with the degrees of freedom equal to the degrees of freedom of the nested model minus the degrees of freedom of the saturated model. If the $LRT_{\text{difference}}$ is not significant, it means that the more parsimonious nested model is not significantly worse than the saturated model. So, then one should choose the nested model since it is simpler (contains less effects).

In some sense, however, the problems associated with Chi-square analysis (see earlier) are also true for the log-linear model. One should therefore be very careful in the interpretation of the results. Again, comparison on the basis of log-linear analysis of the results between different data sets should be avoided.

Empirical Bayes Correction

DuMouchel and Pregibon [2001] and DuMouchel [1999] suggested the use of an Empirical Bayes estimate for the interest value in order to account for the sampling variability in the case of small numbers, which is typically the case when the minimum support threshold specified by the user is small. In that case, slight changes in the absolute value of the support have a large impact on the interest value and this should be corrected for by means of a shrinkage estimate. The procedure goes as follows:

- we have a collection of pairs (n,e), where n is the absolute support of the itemset (above the minimum threshold), and where *e* is the expected absolute support value under the assumption of conditional independence
- for each itemset, *n* is assumed to be drawn from a Poisson distribution with mean $\mu=\lambda*e$.
- however, instead of assuming all λ's to be equal, it is assumed that the λ's are distributed according to a family of prior distributions $\pi(\lambda \mid \theta)$, such as a Gamma distribution or a mixture of Gamma's to have more parameters and thus be more flexible.
- the unconditional distribution for each *n* can now be calculated as $f(n) = \int Poi(n \mid \lambda e)\pi(\lambda \mid \theta)d\lambda$ where *Poi* is the Poisson distribution.
- from the product of this unconditional distribution over all the itemsets, the maximum likelihood estimate $\widehat{\theta}$ can be calculated.
- Using this maximum likelihood estimate $\widehat{\theta}$, we can now calculate the posterior density of λ for each pair (*n*, *e*) as $Poi(n \mid \lambda e)\pi(\lambda \mid \widehat{\theta}) / f(n)$
- the mean of this posterior distribution for each parameter λ will now provide us with an Empirical Bayes shrinkage estimate of the true interest value for each itemset.

The procedure can be easily programmed within the software WinBugs 1.4 and has already been successfully applied on several datasets. On some data sets, however, significant autocorrelations of high-order lags are identified implying the need for rather long chains of iterations which may slow down the calculations significantly.

Nevertheless, empirical results illustrate that the method is indeed able to downsize the interest values for itemsets with low counts. For itemsets with large counts, the Empirical Bayes estimate of the interest does almost not differ from the interest calculated on the raw data. This method is thus clearly preferred over the classical interest measure, especially when the support threshold is being small.

Conclusions

The following suggestions can be formulated based on the analysis of the different interestingness measures discussed in the previous paragraphs:

- Confidence is never the preferred method to compare association rules since it does not account for the baseline frequency of the consequent.
- The lift/interest value corrects for this baseline frequency but when the support threshold is very low, it may be instable due to sampling variability. However, when the data set is very large, even a low percentage support threshold will yield rather large absolute support values. In that case, we do not need to worry too much about sampling variability. A drawback of the interest measure is that it cannot be used to compare itemsets or rules of different size since it tends to overestimate the interestingness for large itemsets.
- Sampling variability can be corrected for by the Empirical Bayes estimate of the interest value. It downsizes the interest when the absolute support of the rule is very low. It generates comparable results to the traditional interest measure when the absolute support is large.
- When association rules need to be compared between data sets of different sizes, the Chi-square test for independence and log-linear analysis are not preferred since they are highly dependent on the dataset size. Both measures tend to overestimate the interestingness of itemsets in large datasets.

Bibliography

[Aggarwal & Yu, 1998] C.C. Aggarwal and P.S. Yu . A New Framework for Item Set Generation. In: Proceedings of the ACM PODS Symposium on Principles of Database Systems, Seattle, Washington (USA), 18-24, 1998.

[Agrawal et al., 1993] R. Agrawal, T. Imielinski and A. Swami A. Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington DC (USA), 207-216, 1993.

[Agrawal & Srikant, 1994] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th Int'l Conference on Very Large Databases (VLDB), Santiago, Chile, 487-499, 1994.

[Agresti, 1996] A. Agresti. An Introduction to Categorical Data Analysis. Wiley Series in Probability and Statistics, 1996.

[Brijs et al., 1999] T. Brijs, G. Swinnen, K. Vanhoof and G. Wets. The use of association rules for product assortment decisions: a case study. In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), August 15-18, 254-260, 1999.

[Brijs et al., 2000] T. Brijs, B. Goethals, G. Swinnen, K. Vanhoof and G. Wets. A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model. In: Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, August 20-23, Boston MA (USA), 300-304, 2000.

[Brin et al., 1998] S. Brin, R. Motwani and C. Silverstein. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. In: Data Mining and Knowledge Discovery, Vol. 2(1), 39-68, 1998.

[DuMouchel, 1999] W. DuMouchel. Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System. In: The American Statistician, Vol. 53, 177-202, 1999.

[DuMouchel & Pregibon, 2001] W. DuMouchel and D. Pregibon. Empirical Bayes Screening for Multi-Item Associations. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco (USA), 67-76, 2001.

[Silverstein et al., 1998] C. Silverstein, R. Motwani and S. Brin. Beyond market baskets: Generalizing association rules to correlations. In: Data Mining and Knowledge Discovery, Vol. 2(1), 39-68, 1998.

[Tan et al., 2002] P-N. Tan, V. Kumar and J. Srivastava, J. Selecting the Right Interestingness Measure for Association Patterns. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton (CA), 32-41, 2002.

Author information

Tom Brijs – Limburgs Universitair Centrum, Research Group Data Analysis and Modelling, Universitaire Campus, gebouw-D, B-3590 Diepenbeek, Belgium; e-mail: tom.brijs@luc.ac.be

Koen Vanhoof - Limburgs Universitair Centrum, Research Group Data Analysis and Modelling, Universitaire Campus, gebouw-D, B-3590 Diepenbeek, Belgium; e-mail: koen.vanhoof@luc.ac.be

Geert Wets - Limburgs Universitair Centrum, Research Group Data Analysis and Modelling, Universitaire Campus, gebouw-D, B-3590 Diepenbeek, Belgium; e-mail: geert.wets@luc.ac.be