Implications of Probabilistic Data Modeling for Mining Association Rules

Michael Hahsler[1], Kurt Hornik[2], and Thomas Reutterer[3]

[1] Department of Information Systems and Operations,
Wirtschaftsuniversität Wien, A-1090 Wien, Austria
[2] Department of Statistics and Mathematics,
Wirtschaftsuniversität Wien, A-1090 Wien, Austria
[3] Department of Retailing and Marketing,
Wirtschaftsuniversität Wien, A-1090 Wien, Austria

Abstract. Mining association rules is an important technique for discovering meaningful patterns in transaction databases. In the current literature, the properties of algorithms to mine association rules are discussed in great detail. We present a simple probabilistic framework for transaction data which can be used to simulate transaction data when no associations are present. We use such data and a real-world grocery database to explore the behavior of confidence and lift, two popular interest measures used for rule mining. The results show that confidence is systematically influenced by the frequency of the items in the left-hand-side of rules and that lift performs poorly to filter random noise in transaction data. The probabilistic data modeling approach presented in this paper not only is a valuable framework to analyze interest measures but also provides a starting point for further research to develop new interest measures which are based on statistical tests and geared towards the specific properties of transaction data.

1 Introduction

Mining association rules (Agrawal et al., 1993) is an important technique for discovering meaningful patterns in transaction databases. An association rule is a rule of the form $X \Rightarrow Y$, where X and Y are two disjoint sets of items (itemsets). The rule means that if we find all items in X in a transaction it is likely that the transaction also contains the items in Y.

A typical application of mining association rules is market basket analysis where point-of-sale data is mined with the goal to discover associations between articles. These associations can offer useful and actionable insights to retail managers for product assortment decisions (Brijs et al., 2004), personalized product recommendations (Lawrence et al., 2001), and for adapting promotional activities (Van den Poel et al., 2004). For web-based systems (e.g., e-shops, digital libraries, search engines) associations found between articles/documents/web pages in transaction log files can even be used to automatically and continuously adapt the user interface by presenting associated items together (Lin et al., 2002).

In Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Magdeburg, March 9-11, 2005, Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2005. Accepted for publication.

Association rules are selected from the set of all possible rules using measures of statistical significance and interestingness. *Support*, the primary measure of significance, is defined as the fraction of transactions in the database which contain all items in a specific rule (Agrawal et al., 1993). That is,

$$\text{supp}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \frac{\text{count}(X \cup Y)}{m}, \tag{1}$$

where $\text{count}(X \cup Y)$ represents the number of transactions which contain all items in X or Y, and m is the number of transactions in the database.

For association rules, a minimum support threshold is used to select the most frequent (and hopefully important) item combinations called *frequent itemsets*. The process of finding these frequent itemsets in a large database is computationally very expensive since it involves searching a lattice which in the worst case grows exponentially in the number of items. In the last decade, research has centered on solving this problem and a variety of algorithms were introduced which render search feasible by exploiting various properties of the lattice (see Goethals and Zaki (2004) as a reference to the currently fastest algorithms).

From the frequent itemsets found, rules are generated using certain measures of interestingness, for which numerous proposals were made in the literature. For association rules, Agrawal et al. (1993) suggest *confidence*. A practical problem is that with support and confidence often too many association rules are produced. In this case, additional interest measures, such as e.g. *lift*, can be used to further filter or rank found rules.

Several authors (e.g., Aggarwal and Yu, 1998) constructed examples to show that in some cases the use of support, confidence and lift can be problematic. Instead of constructing such examples, we will present a simple probabilistic framework for transaction data which is based on independent Bernoulli trials. This framework can be used to simulate data sets which only contain random noise and no associations are present. Using such data and a transaction database from a grocery outlet we will analyze the behavior and problems of the interest measures confidence and lift.

The paper is structured as follows: First, we introduce a probabilistic framework for transaction data. In section 3 we describe the used real-world and simulated data sets. In sections 4 and 5 we analyze the implications of the framework for confidence and lift. We conclude the paper with the main findings and a discussion of directions for further research.

2 A simple probabilistic framework for transaction data

A transaction database consists of a series of transactions, each transaction containing a subset of the available items. We consider transactions which are recorded in a fixed time interval of length t. In Figure 1 an example

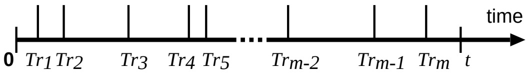

Fig. 1. Transactions occurring over time following a Poisson process.

items

transactions

	l_1	l_2	l_3	...	l_n
p	0.005	0.01	0.0003	...	0.025
Tr_1	0	1	0	...	1
Tr_2	0	1	0	...	1
Tr_3	0	1	0	...	0
Tr_4	0	0	0	...	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Tr_{m-1}	1	0	0	...	1
Tr_m	0	0	1	...	1
c	99	201	7	...	411

Fig. 2. Example transaction database with success probabilities p and transaction counts per item c.

time interval is shown as an arrow with markings at the points in time when the transactions denoted by Tr_1 to Tr_m occur. We assume that transactions occur randomly following a (homogeneous) Poisson process with parameter θ. The number of transactions m in time interval t is then Poisson distributed with parameter θt where θ is the intensity with which transactions occur during the observed time interval:

$$P(M = m) = \frac{e^{-\theta t}(\theta t)^m}{m!} \tag{2}$$

We denote the items which occur in the database by $L = \{l_1, l_2, \ldots, l_n\}$ with n being the number of different items. For the simple framework we assume that all items occur independently of each other and that for each item $l_i \in L$ there exists a fixed probability p_i of being contained in a transaction. Each transaction is then the result of n independent Bernoulli trials, one for each item with success probabilities given by the vector $p = (p_1, p_2, \ldots, p_n)$. Figure 2 contains the typical representation of an example database as a binary incidence matrix with one column for each item. Each row labeled Tr_1 to Tr_m contains a transaction, where a 1 indicates presence and a 0 indicates absence of the corresponding item in the transaction. Additionally, in Figure 2 the success probability for each item is given in the row labeled p and the row labeled c contains the number of transactions each item is contained in (sum of the ones per column).

Following the model, c_i can be interpreted as a realization of a random variable C_i. Under the condition of a fixed number of transactions m this random variable has the following binomial distribution.

$$P(C_i = c_i|M = m) = \binom{m}{c_i} p_i^{c_i}(1 - p_i)^{m-c_i} \tag{3}$$

However, since for a fixed time interval the number of transactions is not fixed, the unconditional distribution gives:

$$\begin{aligned} P(C_i = c_i) &= \sum_{m=c_i}^{\infty} P(C_i = c_i|M = m) \cdot P(M = m) \\ &= \sum_{m=c_i}^{\infty} \binom{m}{c_i} p_i^{c_i}(1 - p_i)^{m-c_i} \, \frac{e^{-\theta t}(\theta t)^m}{m!} \\ &= \frac{e^{-\theta t}(p_i \theta t)^{c_i}}{c_i!} \sum_{m=c_i}^{\infty} \frac{((1 - p_i)\theta t)^{m-c_i}}{(m - c_i)!} \\ &= \frac{e^{-p_i \theta t}(p_i \theta t)^{c_i}}{c_i!} \end{aligned} \tag{4}$$

The sum term in the last but one line in equation 4 is an exponential series with sum $e^{(1-p_i)\theta t}$. With this it is easy to show that the unconditional probability distribution of each C_i has a Poisson distribution with parameter $p_i \theta t$. For short we will use $\lambda_i = p_i \theta t$ and introduce the parameter vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ of the Poisson distributions for all items. This parameter vector can be calculated from the success probability vector p and vice versa by the linear relationship $\lambda = p\theta t$.

For a given database, the values of the parameter θ and the success vectors p or alternatively λ are unknown but can be estimated from the database. The best estimate for θ from a single database is m/t. The simplest estimate for λ is to use the observed counts c_i for each item. However, this is only a very rough estimate which especially gets unreliable for small counts. There exist more sophisticated estimation approaches. For example, DuMouchel and Pregibon (2001) use the assumption that the parameters of the count processes for items in a database are distributed according to a continuous parametric density function. This additional information can improve estimates over using just the observed counts.

3 Simulated and real-world database

We use 1 month ($t = 30$ days) of real-world point-of-sale transaction data from a typical local grocery outlet. For convenience reasons we use categories

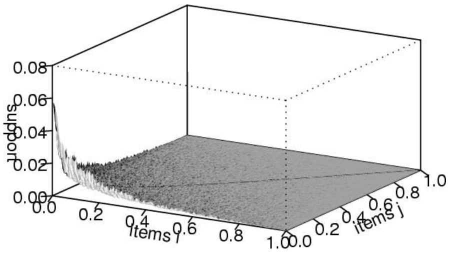

Fig. 3. Support (simulated)

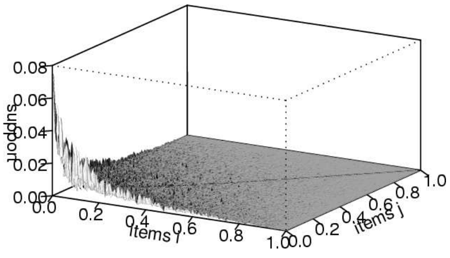

Fig. 4. Support (grocery)

(e.g., *popcorn*) instead of the individual brands. In the available $m = 9835$ transactions we found $n = 169$ different categories for which articles were purchased. The estimated transaction intensity θ for the data set is $m/t = 327.5$ (transactions per day).

We use the same parameters to simulate comparable data using the framework. For simplicity we use the relative observed item frequencies as estimates for λ and calculate the success probability vector p by $\lambda/\theta t$. With this information we simulate the m transactions in the transaction database. Note, that the simulated database does not contain any associations (all items are independent), and thus differs from the grocery database which is expected to contain associations. In the following we will use the simulated data set not to compare it to the real-world data set, but to show that interest measures used for association rules exhibit similar effects on real-world data as on simulated data without any associations.

For the rest of the paper we concentrate on 2-itemsets, i.e., the co-occurrences between two items denoted by l_i and l_j with $i, j = 1, 2, \ldots, n$ and $i \neq j$. Although itemsets and rules of arbitrary length can be analyzed using the framework, we restrict the analysis to 2-itemsets since interest measures for these associations are easily visualized using 3D-plots. In these plots the x and y-axis each represent the items ordered from the most frequent to the least frequent from left to right and front to back and on the z-axis we plot the analyzed measure.

First we compare the 2-itemset support. Figures 3 and 4 show the support distribution of all 2-itemsets. Naturally, the most frequent items also form together the most frequent itemsets (to the left in the front of the plots). The general forms of the two support distributions are very similar. The grocery data set reaches higher support values with a median of 0.000203 compared to 0.000113 for the simulated data. This indicates that the grocery data set contains associated items which co-occur more often than expected under independence.

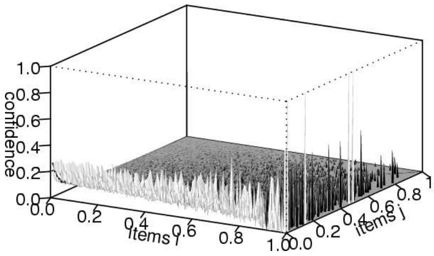

Fig. 5. Confidence (simulated)

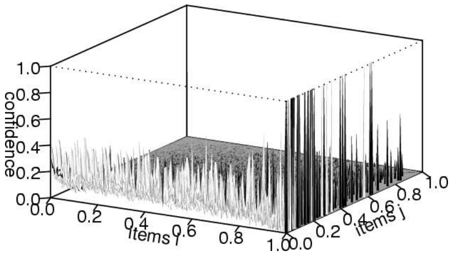

Fig. 6. Confidence (grocery)

4 Implications for the interest measure confidence

Confidence is defined by Agrawal et al. (1993) as

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}, \tag{5}$$

where X and Y are two disjoint itemsets. Often confidence is understood as the conditional probability $P(Y|X)$ (e.g., Hipp et al., 2000), where the definition above is seen as an estimate for this probability.

From the 2-itemsets we generate all rules of the from $l_i \Rightarrow l_j$ and present the confidence distributions in figures 5 and 6. Confidence is generally much lower for the simulated data (with a median of 0.0086 to 0.0140 for the real-world data) which indicates that the confidence measure is able to suppress noise. However, the plots in figures 5 and 6 show that confidence always increases with the item in the right-hand-side of the rule (l_j) getting more frequent. This behavior directly follows from the way confidence is calculated (see equation 5). Especially for the grocery data set in Figure 6 we see that this effect is dominating the confidence measure. The fact that confidence clearly favors some rules makes the measure problematic when it comes to selecting or ranking rules.

5 Implications for the interest measure lift

Typically, rules mined using minimum support (and confidence) are filtered or ordered using their lift value. The measure lift (also called interest, Brin et al., 1997) is defined on rules of the form $X \Rightarrow Y$ as

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)}. \tag{6}$$

A lift value of 1 indicates that the items are co-occurring in the database as expected under independence. Values greater than one indicate that the items are associated. For marketing applications it is generally argued that lift > 1

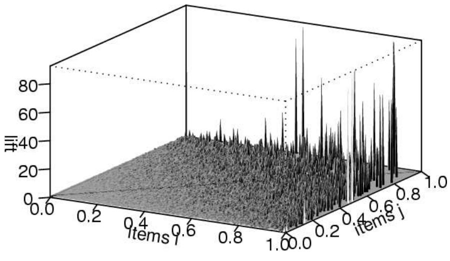

Fig. 7. Lift (simulated)

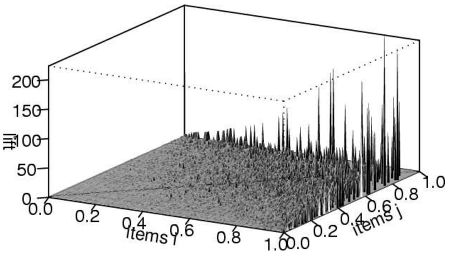

Fig. 8. Lift (grocery)

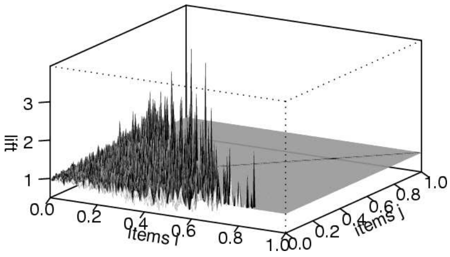

Fig. 9. Lift supp > 0.1% (simulated)

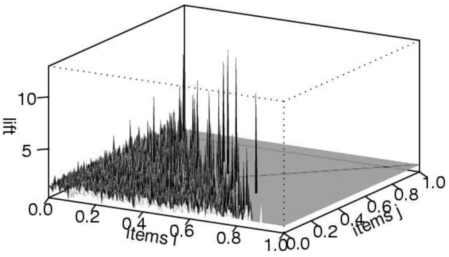

Fig. 10. Lift supp > 0.1% (grocery)

indicates complementary products and lift < 1 indicates substitutes (cf., Betancourt and Gautschi, 1990; Hruschka et al., 1999).

Figures 7 to 10 show the lift values for the two data sets. The general distribution is again very similar. In the plots in Figures 7 and 8 we can only see that very infrequent items produce extremely high lift values. These values are artifacts occurring when two very rare items co-occur once together by chance. Such artifacts are usually avoided in association rule mining by using a minimum support on itemsets. In Figures 9 and 10 we applied a minimum support of 0.1%. The plots show that there exist rules with higher lift values in the grocery data set than in the simulated data. However, in the simulated data we still find 64 rules with a lift greater than 2. This indicates that the lift measure performs poorly to filter random noise in transaction data especially if we are also interested in relatively rare items with low support. The plots in Figures 9 and 10 also clearly show lift's tendency to produce higher values for rules containing less frequent items resulting in that the highest lift values always occur close to the boundary of the selected minimum support. We refer the reader to Bayardo and Agrawal (1999) for a theoretical treatment of this effect. If lift is used to rank discovered rules this means that there is not only a systematic tendency towards favoring rules with less frequent items but the rules with the highest lift will always change with changing the user-specified minimum support.

6 Conclusion

In this contribution we developed a simple probabilistic framework for transaction data based only on independent items. The framework can be used to simulate transaction data which only contains noise and does not include associations. We showed that mining association rules on such simulated transaction data produces similar distributions for interest measures (support, confidence and lift) as on real-world data. This indicates that the framework is appropriate to describe the basic stochastic structure of transaction data.

By comparing the results from the simulated data with the results from the real-world data, we showed how the interest measures are systematically influenced by the frequencies of the items in the corresponding itemsets or rules. In particular, we discovered that the measure lift performs poorly to filter random noise and always produces the highest values for the rules containing the least frequent items. These findings suggest that the existing interest measures need to be supplemented by suitable statistical tests which still need to be developed. Using such tests will improve the quality of the mined rules and the reliability of the mining process.

The presented framework provides many opportunities for further research. For example, explicit modeling of dependencies between items would enable us to simulate transaction data sets with properties close to real data and with known associations. Such a framework would provide an ideal test bed to evaluate and to benchmark the effectiveness of different mining approaches and interest measures. The applicability of the proposed procedure also comprises the development of possible tests against the independence model. Another research direction is to develop new interest measures based on the probabilistic features of the presented framework. A first step in this direction was already done by Hahsler et al. (2005).

Bibliography

AGGARWAL, C. C. and YU, P. S. (1998): A new framework for itemset generation. In: *PODS 98, Symposium on Principles of Database Systems.* Seattle, WA, USA, 18–24.

AGRAWAL, R., IMIELINSKI, T. and SWAMI, A. (1993): Mining association rules between sets of items in large databases. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data.* Washington D.C., 207–216.

BAYARDO, R. J., JR. and AGRAWAL, R. (1999): Mining the most interesting rules. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery in Databases & Data Mining (KDD99).* 145–154.

BETANCOURT, R. and GAUTSCHI, D. (1990): Demand complementarities, household production and retail assortments. *Marketing Science, 9(2), 146–161.*

BRIJS, T., SWINNEN, G., VANHOOF, K. and WETS, G. (2004): Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery, 8(1), 7–23.*

BRIN, S., MOTWANI, R., ULLMAN, J. D. and TSUR, S. (1997): Dynamic itemset counting and implication rules for market basket data. In: *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data.* Tucson, Arizona, USA, 255–264.

DUMOUCHEL, W. and PREGIBON, D. (2001): Empirical Bayes screening for multi-item associations. In: F. Provost and R. Srikant (Eds.): *Proceedings of the ACM SIGKDD Intentional Conference on Knowledge Discovery in Databases & Data Mining (KDD01).* ACM Press, 67–76.

GOETHALS, B. and ZAKI, M. J. (2004): Advances in frequent itemset mining implementations: Report on FIMI'03. *SIGKDD Explorations, 6(1), 109–117.*

HAHSLER, M., HORNIK, K. and REUTTERER, T. (2005): Implications of probabilistic data modeling for rule mining. Report 14, Research Report Series, Department of Statistics and Mathematics, Wirschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria.

HIPP, J., GÜNTZER, U. and NAKHAEIZADEH, G. (2000): Algorithms for association rule mining — A general survey and comparison. *SIGKDD Explorations, 2(2), 1–58.*

HRUSCHKA, H., LUKANOWICZ, M. and BUCHTA, C. (1999): Cross-category sales promotion effects. *Journal of Retailing and Consumer Services, 6(2), 99–105.*

LAWRENCE, R. D., ALMASI, G. S., KOTLYAR, V., VIVEROS, M. S. and DURI, S. (2001): Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery, 5(1/2), 11–32.*

LIN, W., ALVAREZ, S. A. and RUIZ, C. (2002): Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery, 6(1), 83–105.*

VAN DEN POEL, D., DE SCHAMPHELAERE, J. and WETS, G. (2004): Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market. *Expert Systems with Applications, 27(1), 53–62.*