Relative Linkage Disequilibrium: A New Measure for Association Rules

Ron Kenett[1] and Silvia Salini[2]

[1] KPA Ltd., Raanana, Israel and University of Torino, Torino, Italy
ron@kpa.co.il
[2] Department of Economics Business and Statistics, University of Milan, Italy
silvia.salini@unimi.it

Abstract. Association rules are one of the most popular unsupervised data mining methods. Once obtained, the list of association rules extractable from a given dataset is compared in order to evaluate their importance level. The measures commonly used to assess the strength of an association rule are the indexes of support, confidence, and the lift.

Relative Linkage Disequilibrium (RLD) was originally proposed as an approach to analyse both quantitatively and graphically general two way contingency tables. RLD can be considered an adaptation of the lift measure with the advantage that it presents more effectively the deviation of the support of the whole rule from the support expected under independence given the supports of the LHS (A) and the RHS (B). RLD can be interpreted graphically using a simplex representation leading to powerful graphical display of association relationships. Moreover the statistical properties of RLD are known so that confirmatory statistical tests of significance or basic confidence intervals can be applied.

This paper will present the properties of RLD in the context of association rules and provide several application examples to demonstrate it's practical advantages.

Keywords: contingency table, simplex representation, text mining.

1 Introduction

In evaluating the structure of a 2x2 contingency table we consider four relative frequencies, $x_1, x_2, x_3, x_4, \sum_{i=1}^{4} x_i = 1, 0 \leq x_i, i = 1...4$. In the context of this paper, the two variables we consider are occurrence, in a set of transactions, of items A and B on the Left Hand Side and Right Hand Side of an association rule. The frequencies are described in the table below:

In evaluating the association between the two variables, several measures of association are available such as the cross product or odds ratio, $\alpha = \frac{x_1 x_4}{x_2 x_3}$, the chi square statistic [2], the Cramer's index, among others (see [1]). An inherent advantage to informative graphical displays is that the experience and intuition of the experimenter who collects the data can contribute to the statistician's

P. Perner (Ed.): ICDM 2008, LNAI 5077, pp. 189–199, 2008.

	B	^B
A	x_1	x_2
^A	x_3	x_4

data analysis. In [3] a graphical representation of 2x2 tables is proposed using a simplex and, a measure of association, the Relative Linkage Disequilibrium with an intuitive visual interpretation, is proposed. This paper expands on these ideas with a specific focus on association rules that are used, among other things, for analyzing semantic unstructured data.

In Section 2 some details about Relative Linkage Disequilibrium (RLD) and simplex representation are given. Section 3 brievely describes the association rules and the classical measure used to select the rules, moreover the implementation of RLD in the association rules context is described. Section 4 presents two practical application of the RLD; the first example considers a classical market basket analysis data set and compare the RLD with the classical measure; the second example shows the simplex representation and its interesting interpretation using a data set related to the event category (EC) of an electronic product under constant monitoring.

2 Relative Linkage Disequilibrium and Simplex Representation

Relative Linkage Disequilibrium (RLD) is an association measure motivated by indices used in population genetics to assess stability of the genetic composition of populations under various forces of natural selection and migration patterns (see[4] , [5], [6] and [7]). One specific such measure is the linkage disequilibrium, $D = x_1x_4 - x_2x_3$. Under independence, the odds ratio, $\alpha = 1$ and $D = 0$.

There is a natural one to one correspondence between the set of all possible 2x2 contingency tables and point on a simplex. We exploit this graphical representation to map out association rules derived, for example, from text analysis. The tables that correspond to independence in the occurrence of A and B, correspond to a specific surface within the simplex (see figure 1). On that surface, $\alpha = 1$ and $D = 0$.

Let $f = x_1 + x_3$ and $g = x_1 + x_2$. It can be easily verified that:

$x_1 = fg + D$
$x_2 = (1 - f)g - D$
$x_3 = f(1 - g) - D$
$x_4 = (1 - f)(1 - g) + D$

The geometric interpretation of D makes it an appealing measure of interaction. However points closer to the edges of the simplex will have intrinsically smaller values of D. The Relative Linkage Disequilibrium standardizes D by the distance D_M from point corresponding to the contingency table in the simplex to the surface D=0 in an orthogonal projection. RLD is therefore computed as D/D_M.

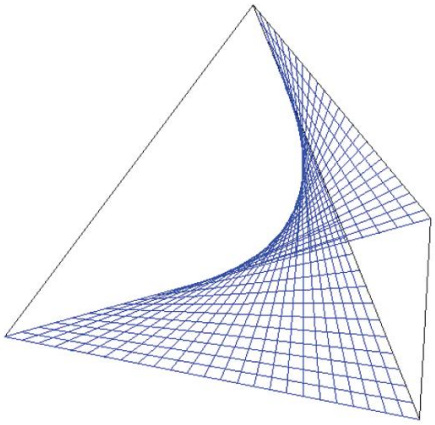

Fig. 1. The surface D=0

The computation of RLD can be performed through the following algorithm:

If $D > 0$
then
    if $x_3 < x_2$
    then $RLD = \frac{D}{D+x_3}$
    else $RLD = \frac{D}{D+x_2}$
else
    if $x_1 < x_4$
    then $RLD = \frac{D}{D-x_1}$
    else $RLD = \frac{D}{D-x_4}$

Some asymptotic properties of RLD are available [3] and can be used for statistical inference.

3 Association Rules and Relative Linkage Disequilibrium

Association rules are one of the most popular unsupervised data mining methods [8]. They were developed in the field of computer science and typically used in applications such as market basket analysis to measure the association between products purchased by each consumer, or in web clickstream analysis, to measure the association between the pages seen (sequentially) by a visitor of a site. Association rules belong to the category of local models, i.e. methods that deal with selected parts of the dataset in the form of subsets of variables or subsets of observations, rather than being applied to the whole database. This element constitutes both the strength and the weak point of the approach. The strength is in that being local, they do not require a large effort from a computational point of view. On the other hand, the locality itself means that a generalization

of the results cannot be allowed, not all the possible relations are evaluated at the same time.

Mining frequent *itemsets* and association rules is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro in [9] describes analyzing and presenting strong rules discovered in databases using different measures of interest. The structure of the data to be analyzed is typically referred to as transactional in a sense explained below.

Let I = $\{i1, i2, \ldots, in\}$ be a set of n binary attributes called "items". Let T = $\{t1, t2, \ldots, tm\}$ be a set of transactions called the database. Each transaction in T has a unique transaction ID and contains a subset of the items in I. Note that each individual can possibly appear more than once in the dataset. In market basket analysis, a transaction means a single visit to the supermarket, for which the list of products bought is recorded, while in web clickstream analysis, a transaction means a web session, for which the list of all visited web-pages is recorded. From this very topic specific structure, the more common data matrix can be easily derived, a different transaction (client) for each row, and a product (page viewed) for each column. The internal cells are filled with 0 or 1 according to the presence or absence of the product (page).

A rule is defined as an implication of the form X => Y where X,Y ∈ I and X ∩Y = ∅;. The sets of items (for short *itemsets*) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule. In an *itemset*, each variable is binary, taking two possible values only, "1" if a specific condition is true, and "0" otherwise.

Each association rule describes a particular local pattern, based on a restricted set of binary variables, and represents relationships between variables which are binary by nature. ,In general, however, this does not have to be the case and continuous rules are also possible. In this case, the elements of the rules can be intervals on the real line, that are conventionally assigned a value of TRUE= 1 and FALSE=0. For instance, a rule of this kind can be X>0 => Y> 100.

Once obtained, the list of association rules extractable from a given dataset is compared in order to evaluate their importance level. The measures commonly used to assess the strength of an association rule are the indexes of support, confidence, and lift.

- The **support** for a rule A => B is obtained by dividing the number of transactions which satisfy the rule, N{A=>B}, by the total number of transactions, N

$$support\ \{A{=>}B\} = N\{A{=>}B\}\ /\ N$$

  The support is therefore the frequency of events for which both the LHS and RHS of the rule hold true. The higher the **support** the stronger the information that both type of events occur together.

- The **confidence** of the rule A => B is obtained by dividing the number of transactions which satisfy the rule N{A=>B} by the number of transactions which contain the body of the rule, A.

$$\textit{confidence}\ \{\text{A=>B}\} = \text{N}\{\text{A=>B}\}\ /\ \text{N}\{\text{A}\}$$

The confidence is the conditional probability of the RHS holding true given that the LHS holds true. A high **confidence** that the LHS event leads to the RHS event **implies causation** or statistical dependence.

– The **lift** of the rule A => B is the deviation of the support of the whole rule from the support expected under independence given the supports of the LHS (A) and the RHS (B).

$$\textit{lift}\ \{\text{A=>B}\} = \textit{confidence}\{\text{A=>B}\}\ /\ \textit{support}\{\text{B}\}$$

$= \textit{support}\{\text{A=>B}\}/\textit{support}\{\text{A}\}\textit{support}\{\text{B}\}$
Lift is an indication of the effect that knowledge that LHS holds true has on the probability of the RHS holding true.

- **when lift is exactly 1:** No effect (LHS and RHS independent). No relationship between events.
- **for lift greater than 1:** Positive effect (given that the LHS holds true, it is more likely that the RHS holds true). Positive dependence between events.
- **if lift is smaller than 1:** Negative effect (when the LHS holds true, it is less likely that the RHS holds true). Negative dependence between events.

Relative Linkage Disequilibrium (RLD) is an alternative measure of association between A and B. For example, take a specific relation A=>B which is observed in 57 cases out of 254 transactions. The LHS, A, is observed without B on the RHS (ˆRHS) in 40 cases, The RHS B is observed in 109 cases without A on the LHS and, in 48 cases, neither A nor B were observed. In our example we obtain the following table:

	B	ˆB
A	57	40
ˆA	109	48

	B	ˆB
A	X_1	X_2
ˆA	X_3	X_4

and in general
Let $x_i = \frac{X_i}{N}, i = 1...4$
For this data:

– $\textit{Support}\{\text{A =>B}\} = 57/254 = .224$
– $\textit{Confidence}\{\text{A =>B}\} = 57/97 = .588$
– $\textit{Support}\ \{\text{B}\} = 166/254 = .654$
– $\textit{Lift}\{\text{A=>B}\} = .588/.654 = .90$
– $D = x_1x_4 - x_2x_3 = (57*48 - 40*109)/(254*254) = -1624/64516 = -0.0252$ and since D<0 and $x_1 > x_4$ we have that
– **RLD**$= D/(D - x4) = -0.0252/(-0.0252 - 0.189) = 0.118$

4 Application Examples

The **arules** extension package for R [10] provides the infrastructure needed to create and manipulate input data sets for the mining algorithms and for analyzing the resulting *itemsets* and rules. Since it is common to work with large sets of rules and *itemsets*, the package uses sparse matrix representations to minimize memory usage. The infrastructure provided by the package was also created to explicitly facilitate extensibility, both for interfacing new algorithms and for adding new types of interest measures and associations.

The library **arules** provides the function interestMeasure() which can be used to calculate a broad variety of interest measures for *itemsets* and rules. All measures are calculated using the quality information available from the sets of *itemsets* or rules (i.e., support, confidence, lift) and, if necessary, missing information is obtained from the transactions used to mine the associations. For example, available measures for *itemsets* are:

- All-confidence [11]
- Cross-support ratio [12]

For rules the following measures are implemented:

- Chi square measure [13]
- Conviction [14]
- Hyper-lift and hyper-confidence[15]
- Leverage [9]
- Improvement [16]
- Several measures from [17] (e.g., cosine, Gini index, ϕ- coefficient, odds ratio)

In this paper we implement the Relative Linkage Disequilibrium measure (RLD) in the function InterestMeasure() and we use the function quadplot() and triplot() of the library **klaR** [18] to produce the simplex 3D and 2D representation.

The first example that we consider is an application to a classical market basket analysis data set. The Groceries (provided by [14] data set contains 1 month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions and the items are aggregated to 169 categories.

In order to compare the classical measure of association rule with RLD we plot in Figure 2 the measures of the 430 rules obtained with the apriori algorithm [19] setting minimum support 0.01 and minimum confidence 0.1.

The plot shows that RLD, like confidence and lift, is able to identify rules that have similar support. Moreover for low level of confidence, the value of RLD is more variable. The relationship with lift is interesting. It seems that RLD can differentiate between groups of rules with the same level of lift.

In Table 1, the first 20 rules sorted by lift are displayed. For each rule the RLD, the odds Ratio and the Chi Square are reported. Figure 3 shows the value of RLD versus odds ratio and versus Chi Square for the top 10 rules.

As we expect for the relationship between RLD and odds ratio (see [1]) the two measure are coherent but different. The Chi Square appears not correlated with

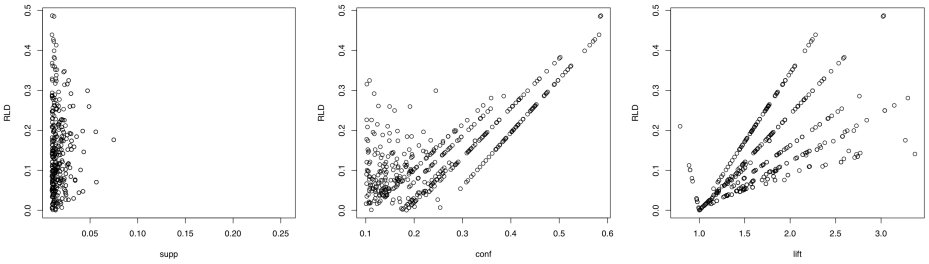

Fig. 2. Plot of Relative Disequilibrium versus Support, Confidence and Lift for the 430 rules of Groceries data set

Table 1. First 20 rules for groceries data, sorted by Lift

lhs	**rhs**	**supp**	**conf**	**lift**	**RLD**	**odds**	**chi**
{whole milk, yogurt}	{curd}	0.010	0.180	3.372	0.141	4.566	184.870
{citrus fruit, other vegetables}	{root vegetables}	0.010	0.359	3.295	0.281	4.958	188.438
{other vegetables,yogurt}	{whipped/sour cream}	0.010	0.234	3.267	0.175	4.450	177.154
other vegetables}	{root vegetables}	0.012	0.343	3.145	0.262	4.679	206.042
{root vegetables}	{beef}	0.017	0.160	3.040	0.250	4.631	277.341
{beef}	{root vegetables}	0.017	0.331	3.040	0.250	4.631	277.341
{citrus fruit, root vegetables}	{other vegetables}	0.010	0.586	3.030	0.487	6.183	175.058
{tropical fruit,, root vegetables}	{other vegetables}	0.012	0.585	3.021	0.485	6.195	207.203
{other vegetables, whole milk}	{root vegetables}	0.023	0.310	2.842	0.225	4.390	330.231
{other vegetables, whole milk}	{butter}	0.012	0.154	2.771	0.143	3.639	146.317
{whole milk, curd}	{yogurt}	0.010	0.385	2.761	0.286	4.088	132.726
{whipped/sour cream}	{curd}	0.011	0.146	2.742	0.135	3.539	129.718
{curd}	{whipped/sour cream}	0.011	0.197	2.742	0.135	3.539	129.718
{other vegetables, whole milk}	{whipped/sour cream}	0.015	0.196	2.729	0.140	3.702	183.728
{other vegetables, yogurt}	{root vegetables}	0.013	0.297	2.729	0.212	3.791	163.187
{whole milk, yogurt}	{whipped/sour cream}	0.011	0.194	2.709	0.132	3.500	131.650
{other vegetables, yogurt}	{tropical fruit}	0.012	0.283	2.701	0.199	3.688	151.333
{root vegetables, other vegetables}	{citrus fruit}	0.010	0.219	2.645	0.148	3.407	119.391
{other vegetables, rolls/buns}	{root vegetables}	0.012	0.286	2.628	0.199	3.568	141.814
{tropical fruit, whole milk}	{root vegetables}	0.012	0.284	2.602	0.196	3.514	136.436

RLD. The major advantage of this new measure is the fact that it is more intuitive than odds ratio and Chi Square and has a useful graphical representation.

In the following example we present the simplex representation and its interesting interpretation. We consider a data set made available by KPA Ltd. The problem consists of mapping the severity level of problems, and the event category (EC) of an electronic product under constant monitoring. Six variables are considered, as shown in Table 2.

The data are recoded as a binary incidence matrix by coercing the data set to transactions. The new data sets present 3733 transactions (rows) and 124 items (columns). Figure 4 shows the item frequency plot (support) of the item with support major than 0.1.

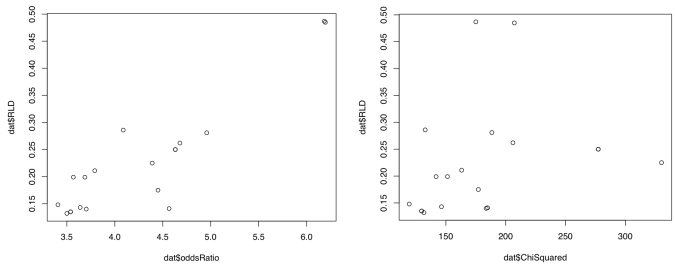

Fig. 3. Plot of Relative Disequilibrium versus Odds Ratio and ChiSquare for the top 10 rules of Groceries data set sorted by RLD

Table 2. Event Category Data Set

PBX No	Severity	Customer Type	EC2	EC1	AL1	AL2	AL3
90009	2	High Tech	SEC08	Security	NO_AL	NP	NP
90009	2	High Tech	NTC09	Network Com	NO_AL	NP	NP
90009	2	High Tech	SEC08	Security	NO_AL	NP	NP
90009	2	High Tech	SEC08	Security	NO_AL	NP	NP
90021	2	Municipalities	SEC08	Security	NO_AL	NP	NP
90033	2	Transportation	SFW05	Software	PCM TS	NP	NP
90033	3	Transportation	INT04	Interface	PCM TS	NP	NP
90033	3	Transportation	SEC05	Security	PCM TS	NP	NP
90038	2	Municipalities	SFW05	Software	NO_AL	NP	NP

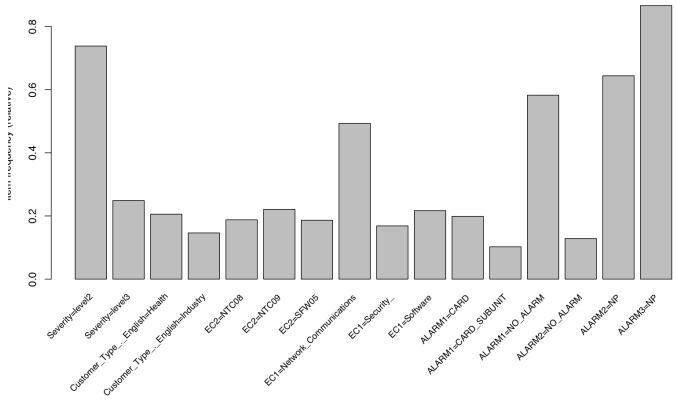

Fig. 4. Item Frequency Plot (Support>0.1) of EC data set

We apply to the data the apriori algorithm setting minimum support 0.1 and minimum confidence 0.8. We obtain 200 rules. The aim of this example is to show the intuitive interpretation of RLD through his useful graphical representation. Figure 5 shows the simplex representation of the contingency tables

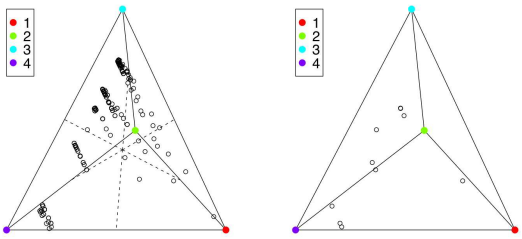

Fig. 5. 3D Simplex Representation for 200 rules of EC data set and for the top 10 rules sorted by RLD

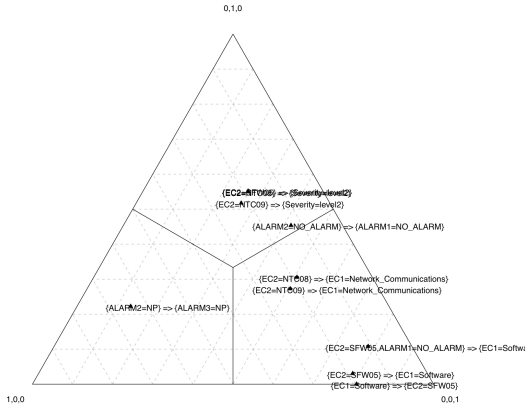

Fig. 6. 2D Simplex Representation for the top 10 rules sorted by RLD of EC data set

corresponding to the rules. The corners represent four tables with relative frequency (1,0,0,0), (01,0,0), (0,0,1,0), (0,0,0,1). We represent the 200 rules obtained from the EC data set and we represent, in the same space, the first 10 rules sorted by RLD.

Figure 5 shows that using a simplex representation, it is possible to immediately have an idea of the rules' structure. In our case, there are 4 groups of rules aligned. Aligned rules imply that they have the same support. In the right part of the Figure 5, the top 10 rules sorted by RLD are plotted. There are some rules in different part of the plot but some of them appear on the same virtual line, only one point is not aligned.

In order to improve the interpretation, we can decide to reduce the dimension and exclude the ˆRHS cell. The 2D representation is shows in Figure 6. In comparison to Table 3, in the left bottom part of the simplex, there are rules with high support, in the right bottom are the rules with low level and in the top are the ones with medium support. The edge in the center represents the middle point of the line, so the point (0.5, 0.5, 0), (0,0.5,0.5) and (0.5, 0, 0.5) obviously assuming ˆRHS equal to 0.

Table 3. Top 10 rules sorted by RLD of EC data set

lhs	rhs	sup	conf	lift	RLD
{EC1=Software}	{EC2=SFW05}	0.1864	0.8593	4.6086	1.0000
{AL2=NO_AL}	{AL1=NO_AL}	0.1286	1.0000	1.7171	1.0000
{EC2=SFW05}	{EC1=Software}	0.1864	1.0000	4.6086	1.0000
{EC2=SFW05}	{Severity=level2}	0.1864	1.0000	1.3550	1.0000
{EC2=NTC08}	{EC1=Network_Com}	0.1878	1.0000	2.0277	1.0000
{EC2=NTC08}	{Severity=level2}	0.1878	1.0000	1.3550	1.0000
{EC2=NTC09}	{EC1=Network_Com}	0.2207	1.0000	2.0277	1.0000
{EC2=NTC09}	{Severity=level2}	0.2207	1.0000	1.3550	1.0000
{AL2=NP}	{AL3=NP}	0.6440	1.0000	1.1543	1.0000
{EC2=SFW05,AL1=NO_AL}	{EC1=Software}	0.1090	1.0000	4.6086	1.0000

5 Summary and Future Work

Relative Linkage Disequilibrium is a useful measure in the context of association rules, especially for its intuitive visual interpretation. An inherent advantage to informative graphical displays is that the experience and intuition of the experimenter who collects the data can contribute to the statistician's data analysis.

The examples proposed in this paper show that RLD, like confidence and lift, is able to identify rules that have similar support. Moreover for low level of confidence, the value of RLD is more variable. The relationship with lift is interesting, it seems that RLD can differentiate between groups of rules with the same level of lift. Moreover RLD is coherent with odds ratio and differ from Chi Square. The second example highlight the major advantage of the new measure: it is more intuitive than odds ratio and Chi Square and has a useful graphical representation that make possible to immediately have an idea of the rules' structure and to identify groups of rules.

In our future intention RLD can be applied to text mining and web mining data and to the analysis of comments in survey questionnaires. The context of application of RLD ranges from cognitive science, ability tests and customer satisfaction surveys, in order to find associated item and to test if there are some redundant items.

Another important future work is the exploration of statistical properties of RLD in the context of association rules. For a partial study of such properties see [3].

References

1. Kenett, R., Zacks, S.: Modern Industrial Statistics. Duxbury Press, San Francisco (1998)
2. Shimada, K., Hirasawa, K., Hu, J.: Association Rule Mining with Chi-Squared Test Using Alternate Genetic Network Programming. In: Perner, P. (ed.) ICDM 2006. LNCS (LNAI), vol. 4065, pp. 202–216. Springer, Heidelberg (2006)

3. Kenett, R.: On an Exploratory Analysis of Contingency Tables. The Statistician 32, 395–403 (1983)
4. Fisher, R.A.: The Genetical Theory of Natural Selection. Clarendon, Oxford (1930)
5. Lewontin, R., Kojima, K.: The evolutionary dynamics of complex polymorphisms. Evolution 14, 458–472 (1960)
6. Karlin, S., Feldman, M.: Linkage and selection: Two locus symmetric viability model. Theoretical Population Biology 1, 39–71 (1970)
7. Karlin, S., Kenett, R.: Variable Spatial Selection with Two Stages of Migration and Comparisons Between Different Timings. Theoretical Population Biology 11, 386–409 (1977)
8. Agrawal, R., Imielienski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proc. Conf. on Management of Data, pp. 207–216. ACM Press, New York (1993)
9. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Knowledge Discovery in Databases, pp. 229–248 (1991)
10. Hahsler, M., Gr¨un, B., Hornik, K.: arules – A computational environment for mining association rules and frequent item sets. Journal of Statistical Software 14(15), 1–25 (2005), http://www.jstatsoft.org/v14/i15/
11. Omiecinski, E.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15(1), 57–69 (2003)
12. Xiong, H., Tan, P.-N., Kumar, V.: Mining strong affinity association patterns in data sets with skewed support distribution. In: Goethals, B., Zaki, M.J. (eds.) Proceedings of the IEEE International Conference on Data Mining, Melbourne, Florida, November 19–22, pp. 387–394 (2003)
13. Liu, B., Hsu, W., Ma, Y.: Pruning and summarizing the discovered associations. In: KDD 1999: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 125–134. ACM Press, New York (1999)
14. Brin, S., Motwani, R., Ullman, J., Tsur, S.: 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, pp. 255–264 (1997)
15. Hahsler, M., Hornik, K., Reutterer, T.: Implications of probabilistic data modeling for mining association rules. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nuernberger, A., Gaul, W. (eds.) From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 598–605. Springer, Heidelberg (2006)
16. Bayardo, R., Agrawal, R., Gunopulos, D.: Constraint-based rule mining in large, dense databases. Data Mining and Knowledge Discovery 4(2/3), 217–240 (2000)
17. Tan, P.-N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4), 293–313 (2004)
18. Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., Zentgraf, M. (2008), http://www.statistik.tu-dortmund.de
19. Borgelt, C.: Apriori – Finding Association Rules/Hyperedges with the Apriori Algorithm. Working Group Neural Networks and Fuzzy Systems, Otto-von-Guericke-University of Magdeburg, Universit ¨atsplatz 2, D-39106 Magdeburg, Germany (2004), http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html