Chapter 9

Measures of Association Applied to Operational Risks

R. Kenett and S. Salini

Synopsis

Association rules are a basic analysis tools for unstructured data such as accident reports, call centres recordings and CRM logs. Such tools are commonly used in basket analysis of shopping carts for identifying patterns in consumer behaviour. In this chapter we show how association rules are used to analyze unstructured operational risk data in order to provide risk assessments and diagnostic insights. We will also present a new graphical display of association rules that permits effective clustering of associations with a novel association rule called the Relative Linkage Disequilibrium.

Keywords: Association rules, Data Mining, Relative Linkage Disequilibrium (RLD), itemsets, interest measures.

9.1 Introduction

Association rules are one of the most popular unsupervised data mining methods (Agarwal et al, 1993, Borgelt et al, 2004, Kenett and Salini, 2008a and 2008b, Roever et al, 2008 and Tan et al, 2004). 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 consumers, or in web clickstream analysis, to measure the association between the pages seen by a visitor of a site. Sequence rules algorithms are employed to analyse also the sequence of pages seen by a visitor.

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, 1991, 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 $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called "items". Let $T = \{t_1, t_2, \ldots, t_m\}$ 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 the continuous 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 example, 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.

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

  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).

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

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

  Lift is an indication of the effect that knowledge that LHS holds true has on the probability of the RHS holding true. Hence Lift is a value that gives us information about the increase in probability of the "then" (consequent RHS) given the "if" (antecedent LHS) part.
  - 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 association measure motivated by indices used in population genetics to assess stability over time in the genetic composition of populations. This same measure has been also suggested as an exploratory analysis methods applied to general 2x2 contingency tables (see Kenett, 1983 and Kenett and Zacks, 1998). To define RLD, consider a transactions set with item A on the Left Hand Side (LHS) and item B on the Right Hand Side (RHS) of an association rule. In a specific set of transactions, these two events generate four combinations whose frequencies are described in Table 9.1 below:

Table 9.1: The association rules contingency table of A and B

	B	^B
A	x_1	x_2
^A	x_3	x_4

$$\sum_{i=1}^{4} x_i = 1,\ 0 \le x_i, i = 1...4.$$

x_1 = the relative frequency of occurrence of both A and B

x_2 = the relative frequency of transactions where only A occurs

x_3 = the relative frequency of transactions where only B occurs

x_4 = the relative frequency of transaction where neither A or B occur

There is a natural one to one correspondence between the set of all possible 2x2 contingency tables, such as Table 9.1, and points on a simplex (see Figure 9.1). We exploit this graphical representation to map out association rules. The tables that correspond to independence in the occurrence of A and B, correspond to a specific surface within the simplex presented in Figure 9.1. By "independence" we mean that knowledge of marginal frequencies of A and B is sufficient to reconstruct the entire table, i.e. the items A and B do not interact.

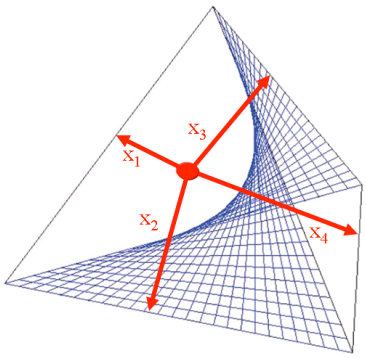

Figure nce (D=0)

Let $D = x_1x_4 - x_2x_3$ $f = x_1 + x_3$ and $g = x_1 + x_2$.

f = relative frequency of item B

g = relative frequency of item A

The surface in Figure 9.1 corresponds to contingency tables with D=0 (or *lift* = 1).

It can be easily verified that:

$$x_1 = f\,g + D = support\ \{A=>B\}$$
$$x_2 = (1-f)g - D$$
$$x_3 = f(1-g) - D$$
$$x_4 = (1-f)(1-g) + D$$

and that

$$confidence\ \{A=>B\} = \frac{x_1}{x_1 + x_2} = \frac{x_1}{g}$$

$$lift\ \{A=>B\} = \frac{x_1}{(x_1 + x_2)\cdot(x_1 + x_3)} = \frac{x_1}{f\cdot g} = 1 + \frac{D}{f\cdot g},\ (-1 \le D \le 1)$$

The geometric interpretation of D makes it an appealing measure of interaction. As mentioned, the surface on Figure 9.1 represents all association rules with $D = 0$. However points closer to the edges of the simplex will have intrinsically smaller values of D.

Let D_M be the distance from the point corresponding to the contingency table on the simplex to the surface D=0 in the direction (1, -1, -1, 1).

We define Relative Linkage Disequilibrium (RLD) = D/ D_M.

As can be seen geometrically, RLD standardizes D by the maximal distance D_M .

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 $RLD = \frac{D}{D - x_4}$

Asymptotic properties of RLD are available in Kenett, 1983, and RLD can be also used for statistical inference.

9.2 The arules R script library

The **arules** extension package for R (Hashler et al 2005 and 2008) 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 (Omniecinki, 2003)
- Cross-support ratio (Xiong et al, 2003)

For rules the following measures are implemented:

- Chi square measure (Kenett and Zacks, 1998)
- Conviction (Brin et al, 1997)
- Hyper-lift and hyper-confidence (Hashler et al, 2006)
- Leverage (Piatetsky-Shapiro, 1991)
- Improvement (Bayardo et al, 2000)
- Several measures from Tan, 2004, (e.g., cosine, Gini index, φ- coefficient, odds ratio)
- Relative Linkage Disequilibrium (RLD), Kenett and Salini 2008a and 2008b.

As mentioned above, the Relative Linkage Disequilibrium measure (RLD) is in the function "InterestMeasure()". We use the function quadplot() and triplot() of the library **klaR** (Roever, 2008) to produce the simplex 3D and 2D representation.

9.3 Some Examples

9.3.1 Market Basket Analysis

The first example that we consider is an application to a classical market basket analysis data set. The Groceries data set contains 1 month (30 days) of real world point of sale transaction data from a typical local grocery outlet (Hashler, 2008). The data set contains 9835 transactions and the items are aggregated into 169 categories.

In order to compare the classical measure of association rule with RLD, we plot in Figure 9.2 measures of the 430 rules obtained with the a-priori algorithm setting minimum support equal to 0.01 and minimum confidence to 0.1.

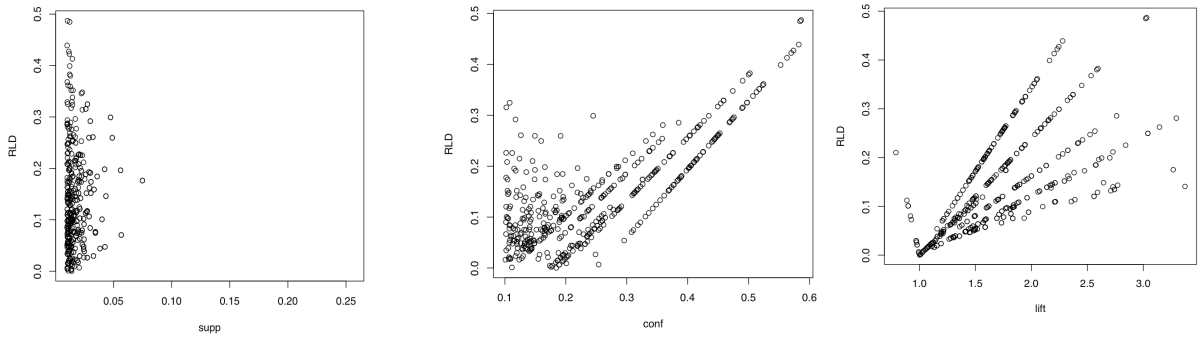

Figure 9.2: Plot of Relative Linkage Disequilibrium versus Support, Confidence and Lift for the 430 rules of Groceries data set

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

Table 9.2 displays the first 20 rules sorted by lift. For each rule, the RLD, the Odds Ratio and the Chi Square values are reported. Figure 9.3 shows the value of RLD versus Odds Ratio and versus Chi Square for the top 10 rules.

Table 9.2: First 20 rules for groceries data, sorted by Lift.

	lhs		rhs	support	confidence	lift	RLD	oddsRatio	chiSquare
1	{whole milk, yogurt}	=>	{curd}	0.01006609	0.1796733	3.372304	0.1407948	4.565544	184.8700
2	{citrus fruit, other vegetables}	=>	{root vegetables}	0.01037112	0.3591549	3.295045	0.2807587	4.957868	188.4380
3	{other vegetables, yogurt}	=>	{whipped/sour cream}	0.01016777	0.2341920	3.267062	0.1750579	4.449668	177.1536
4	{tropical fruit, other vegetables}	=>	{root vegetables}	0.01230300	0.3427762	3.144780	0.2623764	4.678610	206.0424
5	{root vegetables}	=>	{beef}	0.01738688	0.1595149	3.040367	0.2496032	4.630855	277.3405
6	{beef}	=>	{root vegetables}	0.01738688	0.3313953	3.040367	0.2496032	4.630855	277.3405
7	{citrus fruit, root vegetables}	=>	{other vegetables}	0.01037112	0.5862069	3.029608	0.4869320	6.182676	175.0581
8	{tropical fruit, root vegetables}	=>	{other vegetables}	0.01230300	0.5845411	3.020999	0.4848665	6.194803	207.2034
9	{other vegetables, whole milk}	=>	{root vegetables}	0.02318251	0.3097826	2.842082	0.2253466	4.389810	330.2314
10	{other vegetables, whole milk}	=>	{butter}	0.01148958	0.1535326	2.770630	0.1432227	3.638945	146.3170
11	{whole milk, curd}	=>	{yogurt}	0.01006609	0.3852140	2.761356	0.2855465	4.087797	132.7261
12	{whipped/sour cream}	=>	{curd}	0.01047280	0.1460993	2.742150	0.1345253	3.539378	129.7175
13	{curd}	=>	{whipped/sour cream}	0.01047280	0.1965649	2.742150	0.1345253	3.539378	129.7175
14	{other vegetables, whole milk}	=>	{whipped/sour cream}	0.01464159	0.1956522	2.729417	0.1398891	3.701980	183.7284
15	{other vegetables, yogurt}	=>	{root vegetables}	0.01291307	0.2974239	2.728698	0.2114760	3.791185	163.1868
16	{whole milk, yogurt}	=>	{whipped/sour cream}	0.01087951	0.1941924	2.709053	0.1319696	3.500414	131.6497
17	{other vegetables, yogurt}	=>	{tropical fruit}	0.01230300	0.2833724	2.700550	0.1993601	3.688172	151.3326
18	{root vegetables, other vegetables}	=>	{citrus fruit}	0.01037112	0.2188841	2.644626	0.1484010	3.407110	119.3908
19	{other vegetables, rolls/buns}	=>	{root vegetables}	0.01220132	0.2863962	2.627525	0.1990992	3.568197	141.8140
20	{tropical fruit, whole milk}	=>	{root vegetables}	0.01199797	0.2836538	2.602365	0.1960214	3.513535	136.4357

As we expect for the relationship between RLD and Odds Ratio, the two measures are coherent but still different. The Chi Square values appear not to be correlated with RLD so that the information provided by RLD is not redundant with Chi Square. Moreover, RLD is more intuitive than the Odds Ratio and Chi Square since it has a useful graphical interpretation.

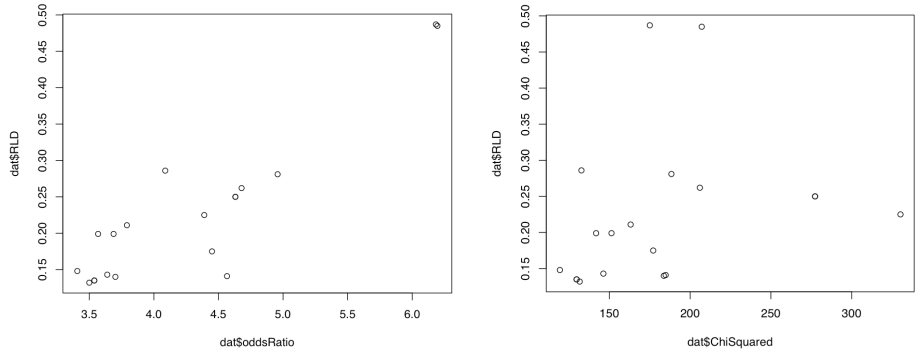

Figure 9.3: Plot of Relative Linkage Disequilibrium versus Odds Ratio and Chi Square for the top 10 rules of Groceries data set sorted by RLD

9.3.2 PBX System Risk Analysis

In the following example we present an analysis of data collected from Private Branch Exchange (PBX) Telecommunication Systems (See also Chapter 5 and Cerchiello and Giudici, 2007).

Operational risks, in this context, are typically classified into hardware, software, interface, network and security related events (See Chapter 3). Assessing operational risks involves merging data from different sources such as system logs, call centre records, technical service data bases and customer complaints (see Chapter 5).

The problem consists of mapping the severity level of problems, and the Event Category (EC) of a PBX under constant monitoring. Seven variables are considered, as shown in Table 9.3. For more details about the data see Cerchiello and Bonafede, 2009.

Table 9.3: Event Category data set

PBX No	Severity	Customer Type	EC2	EC1	ALARM1	ALARM2	ALARM3
90009	2	High Tech	SEC08	Security	NO_ALARM	NP	NP
90009	2	High Tech	NTC09	Network Communications	NO_ALARM	NP	NP
90009	2	High Tech	SEC08	Security	NO_ALARM	NP	NP
90009	2	High Tech	SEC08	Security	NO_ALARM	NP	NP
90021	2	Municipalities	SEC08	Security	NO_ALARM	NP	NP
90033	2	Transportation	SFW05	Software	PCM TIME SLOT	NP	NP
90033	3	Transportation	INT04	Interface	PCM TIME SLOT	NP	NP
90033	3	Transportation	SEC05	Security	PCM TIME SLOT	NP	NP
90038	2	Municipalities	SFW05	Software	NO_ALARM	NP	NP

The data is 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 9.4 shows the item frequency plot (support) of the item with support major than 0.1.

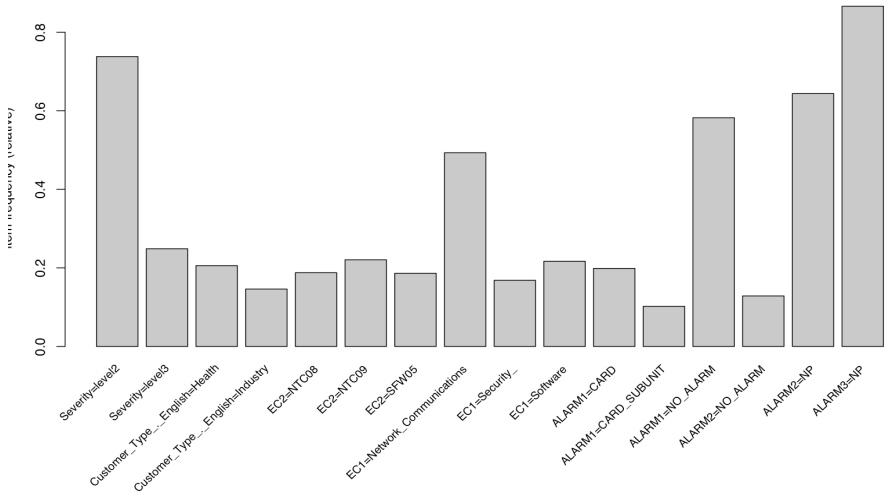

Figure 9.4: Item Frequency Plot (Support>0.1) of EC data set

We apply to the data the apriori algorithm setting minimum support to 0.1 and minimum confidence to 0.8 and obtain 200 rules. The aim of this example is to show the intuitive interpretation of RLD through its useful graphical representation. Figure 9.5 shows the simplex representation of the contingency tables corresponding to these 200 rules. The corners represent tables with relative frequency (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1). The dots on the left figure represent all the rules derived from the EC data set and the dots on the right figure correspond to the first 10 rules sorted by RLD.

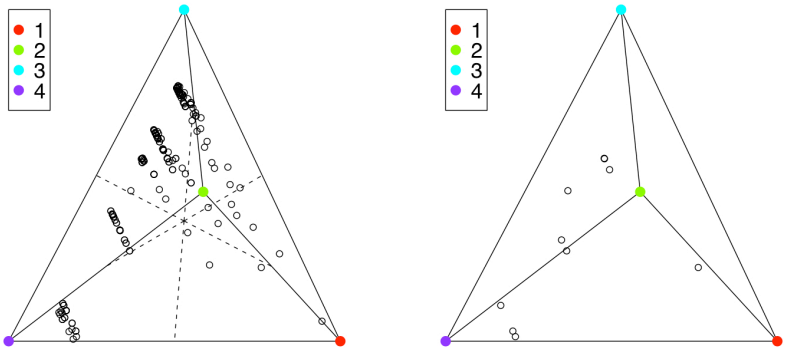

Figure 9.5: 3D Simplex representation for 200 rules of EC data set (left) and for the top 10 rules sorted by RLD (right)

Figure 9.5 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 aligned rules. Aligned rules imply that they have the same support.

In order to improve the interpretation, we can try to reduce the dimensionality of the 2X2 table. A two dimensional representation is shown in Figure 9.6. On the left bottom part of the simplex, there

are rules with high support, on the right bottom there are rules with low support and at the top are the rules with medium support. The table corresponding to the center point is (0,5, 0,5, 0,5, 0,5).

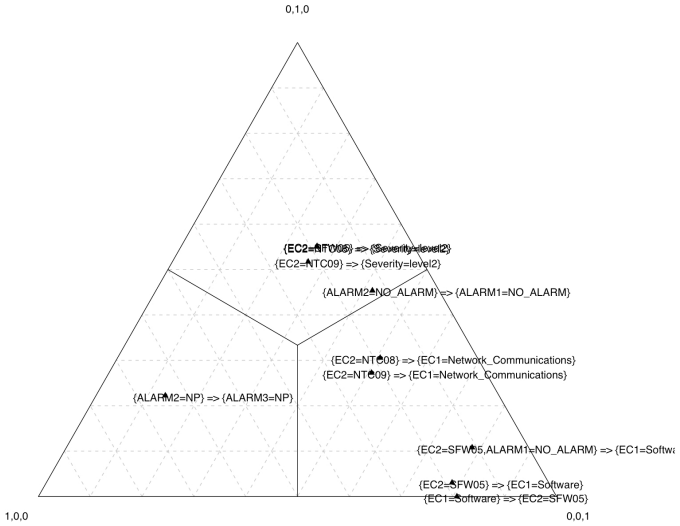

Figure 9.6: 2D Simplex representation for the top 10 rules, sorted by RLD

9.3.3 A Bank's Operational Risk Analysis

Operational risk in the banking industry is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events (Basel, 2004).

These include:

- □ Internal fraud
- □ External fraud
- □ Employment practices & workplace safety
- □ Clients, products & business practices
- □ Damage to physical assets
- □ Business disruption & system failures
- □ Execution, delivery & process management
- □ Includes legal risk.

Operational risks exclude reputational and business/strategic risk.

The rising interest of the banking industry in operational risks is due, among other reasons, to the globalization of the financial markets, the growth of IT applications, and the increasing diffusion of sophisticated financial products. The Basel II capital accord requires banks to put aside a minimum capital requirement which matches its exposure to credit risk, market risk and operational risk. Specifically, a 12% of minimum capital requirement needs to be allocated to operational risks (Basel, 2004).

The Basel II agreement splits operational risk exposures and losses into a series of standardized business units, called '*business lines*', and into groups of operational risk losses according to the nature of the underlying operational risk event, called '*event types*'. In (Basel, 2008) a comprehensive Loss Data Collection Exercise (LDCE) initiated by the Basel II Committee, through the work of its Operational Risk Subgroup of the Accord Implementation Group (AIGOR), is described. The exercise follows other similar exercises sponsored by the Basel Committee and individual member countries over the last five years. The 2008 LDCE is a significant step forward in the Basel Committee's efforts to address Basel II implementation and post-implementation issues

more consistently across member jurisdictions. While similar to two previous international LDCEs, which focused on internal loss data, this LDCE is the first international effort to collect information on all four operational risk data elements: 1) internal data, 2) external data, 3) scenario analysis, and 4) business environment and internal control factors (BEICFs). The BEICFs are used in an Advanced Measurement Approach (AMA) for calculating operational risk capital charges under Basel II. As an independent contribution to the LDCE we present here the application of RLD to internal operational risk data collected by a large banking institution. Our goal is to demonstrate, with a concrete example, how RLD can be used to assess risks reported in such organizations using textual reports.

We consider a data set of operational risk events with 20 variables, some categorical, some continuous and some textual, with a description of the loss event. Examples of such descriptions are:

- "Booked on fixed income trade that was in the wrong pat fund code. Have cancelled trade resultant in error of 15000"
- "Cash contribution not invested due to incorrect fax number used by client. Not our error but noted due to performance impact on the fund."
- "The client sent a disinvestment instruction that was incorrectly processed as an investment. Due to a positive movement in the equity markets the correction of the error led to a gain."

In the data preparation phase, we discretized the continuous variables (expected and actual values of loss) and, using the library **tm** of R (Feinerer, 2007), we selected the textual description variables, in particular, *activity*, *process* and *risk* type. Than, the data was processed for an association rules analysis.

Following these steps, we obtain a new data set with 2515 transactions and 235 items (the levels of the variables). The a-priori algorithm produces 345,575 rules[1]. With such a large number of rules traditional measures of association typically cannot identify "interesting" associations. Too many rules with too little a difference between them. Moreover, with traditional measures of association, it is often difficult to explore and cluster rules in an association rules analysis. RLD and its complementary simplex representation help us in tackling this problem.

For each rule, we calculate RLD and sort the rules accordingly. Figure 9.7 shows the first 200 rules with the highest level of RLD.

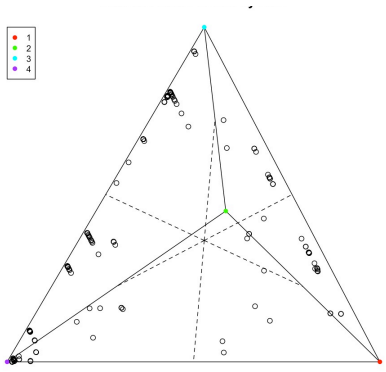

Figure 9.7: Simplex representation of the first 200 rules sorted by RLD for operational risk data set

[1] We modify the default level of support in the arules alghoritm of R, we set a very low level of support 0,01. This is useful in operational risk application, because we expect that the loss event are not so frequent.

We compare the top 200 rules derived from sorting association rules by support, confidence and lift with RLD (see Figure 9.8). RLD clearly provides the highest resolution and interesting spread.

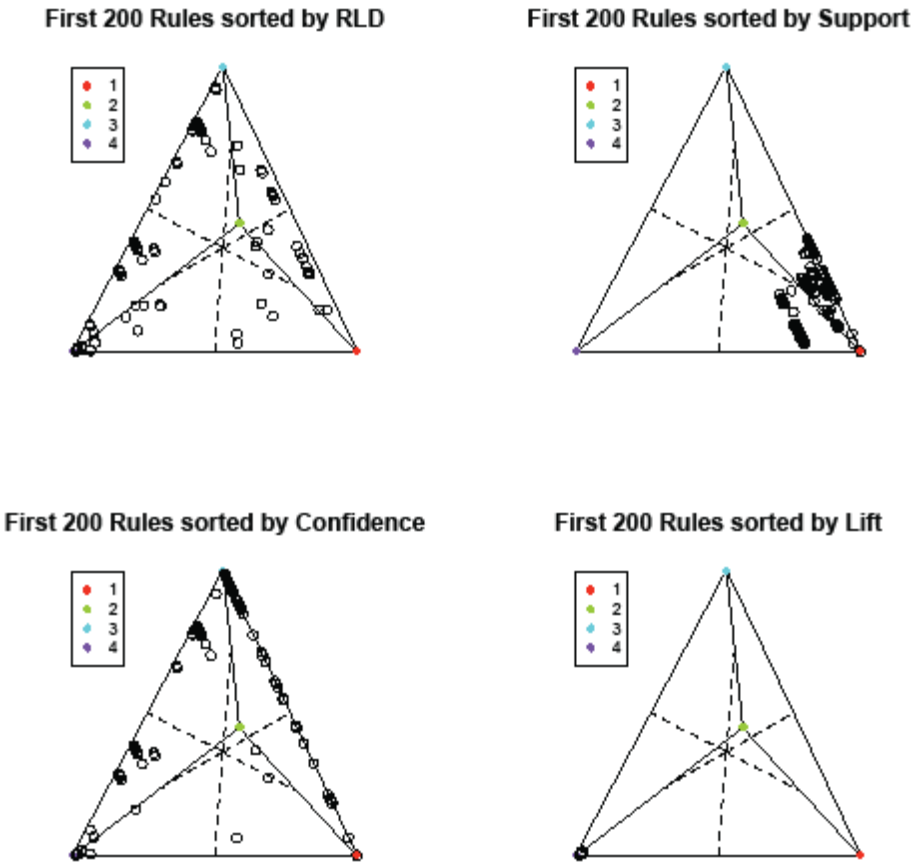

Figure 9.8: Comparison of the first 200 rules sorted by RLD, support, confidence and lift for the operational risk data set

We proceed with an automatic clustering of the rules. This is applied here to the first 200 rules sorted by RLD, but can also be done for other rules.

The hierarchical cluster analysis is applied to the elements in the association rules contingency table on the numbers that we use in the calculation of RLD. Figure 9.9 shows the cluster dendrogram with a highlight of 12 clusters of association rules.

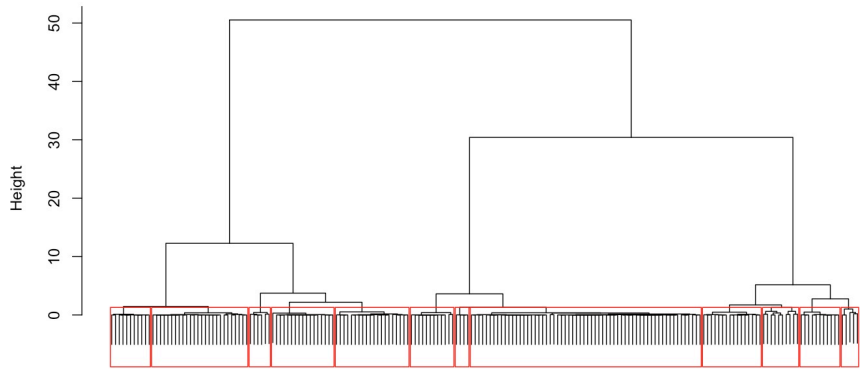

Figure 9.9: Cluster dendrogram for the 200 rules for operational risk data set

Now we produce a simplex representation for each one of the clusters. Figure 9.10 shows these plots. Rules in the same cluster have similar type of association. All the rules in these plots have a very high level of RLD, near 1, but different values for the other association measures. For example the rules in the left bottom corner of the clusters 5, 10 and 12 are characterized by very low support and very high lift. On the contrary rules in clusters 2 and 3 have high support, high confidence and low lift. In cluster 11 there are rules with confidence equal to 1, lift nearer 1 and very low support, etc…

This example demonstrates the unique property of RLD, using a real data set. We conclude with a summary and some direction for future work.

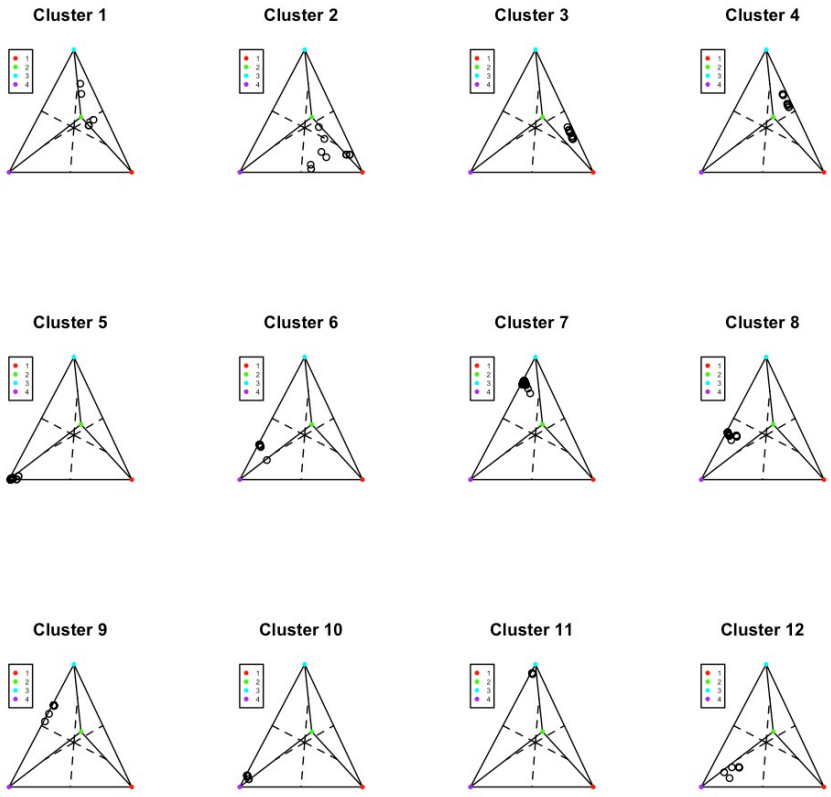

Figure 9.10: Cluster simplex plot for the 200 rules for operational risk data set

9.4 Summary

Relative Linkage Disequilibrium is a useful measure in the context of association rules, especially for its intuitive quantitative and 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. This is an essential component of Information Quality (InfoQ) discussed in Chapter 1.

The context for applications of RLD ranges from web sites logs, customer satisfaction surveys, operational risks data, call centers records and many other sources of textual data. The first two examples presented in this chapter show that RLD, like confidence and lift, is able to identify rules that have similar support. Moreover for low levels of confidence, the value of RLD is more informative. The relationship with lift is interesting, it seems that RLD can differentiate between groups of rules with the same level of lift. RLD is correlated with the Odds Ratio but differs from the Chi Square values. The second example highlights the major advantage of the new measure: it is more intuitive than the Odds Ratio and Chi Square and has a useful graphical representation of the rules' structure and allows us to identify groups of rules. The third example shows how RLD can be used to select and cluster association rules.

RLD can contribute to identify rare events in large text files, events called "Black Swans" (see Chapter 1, Chapter 14 and Taleb, 2007). Combining RLD with simplex representations can help display item sets with low support exhibiting significant association patterns. This chapter provides an introduction to Relative Linkage Disequilibrium with applications to Operational Risk Management. Hopefully it will stimulate more research on association rules and their close relationship with contingency tables.

References

1. Agrawal, R., Imielienski, T., and Swami, A., Mining Association Rules between Sets of Items in Large Databases. *Proc. Conf. on Management of Data*, pp. 207–216. ACM Press, New York (1993).
2. Basel Committee on Banking Supervision (2004). *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework*, www.bis.org/publ/bcbs107.htm.
3. Basel Committee on Banking Supervision (2008). *Operational Risk - 2008 Loss Data Collection Exercise*, www.bis.org/publ/bcbs_nl13.htm.
4. Bayardo, R., Agrawal, R., and Gunopulos, D., Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4, 2/3, pp.217–240 (2000).
5. 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, URL http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html (2004).
6. Brin, S., Motwani, R., Ullman, J., and Tsur, S., Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 255–264, Tucson, Arizona (1997).
7. Cerchiello P., Bonafede E., A Proposal to Fuzzify Categorical Variables in Operational Risk Management. In *Data Analysis and Classification*, Springer, forthcoming (2009).
8. Cerchiello P., Giudici P., Causal Risk Models: A proposal based on associative classes, *Proceedings of the 2007 Intermediate Conference "Risk and Prediction"*, SIS -Società Italiana di Statistica, Venice, pp 545-546 (2007).
9. Feinerer I (2007). *tm: Text Mining Package. R package version 0.3*, URL http://CRAN.R-project.org/package=tm.
10. Hahsler, M., Gr¨un, B., and Hornik, K.,*The arules package: Mining Association Rules and Frequent Itemsets, version 0.6-6,* http://cran.r-project.org/web/packages/arules/index.html (2008).
11. Hahsler, M., Gr¨un, B., and Hornik, K., arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software,* 14(15):1–25. ISSN 1548-7660. URL http://www.jstatsoft.org/v14/i15 (2005).
12. Hahsler, M., Hornik, K., and Reutterer, T., Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 598–605, Springer-Verlag (2006).
13. Kenett, R.S., On an Exploratory Analysis of Contingency Tables. *The Statistician*, 32, pp. 395-403 (1983).
14. Kenett, R.S. and Zacks, S. (1998) *Modern Industrial Statistics: Design and Control of Quality and Reliability*, Duxbury Press, San Francisco, 1998, Spanish edition 2000, 2nd paperback edition 2002, Chinese edition (2004).
15. Kenett, R.S. and Salini, S., Relative Linkage Disequilibrium: A new measure for association rules", in: P. Perner (Ed.), *Advances in Data Mining: Medial Applications, E-Commerce, Marketing, and Theoretical Aspects*, ICDM 2008, Leipzig, Germany, July, 2008. Lecture Notes in Computer Science, Springer Verlag, Vol. 5077 (2008a).
16. Kenett, R.S. and Salini, S., Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks, *Transactions on Machine Learning and Data Mining*, Vol.1, No 2, pp. 83-96 (2008b).
17. MUSING - MUlti-industry, Semantic-based next generation business INtelliGence (IST- FP6 27097) http://www.musing.eu (2006).
18. Omiecinski, E., Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15, 1, pp. 57–69 (2003).

19. Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pp. 229–248 (1991).
20. Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., Zentgraf , M., *The klaR package: Classification and visualization, version 0.5-7*, http://cran.r-project.org/web/packages/klaR/index.html (2008).
21. Taleb, N. (2007). *The Black Swan: The impact of the highly improbable*, Random House, NY.
22. Tan, P.-N., Kumar, V., and Srivastava, J., Selecting the right objective measure for association analysis. *Information Systems*, 29, 4, pp.293–313 (2004).
23. Xiong H., Tan P.-N., and Kumar V., Mining strong affinity association patterns in data sets with skewed support distribution. In B. Goethals and M. J. Zaki, editors, *Proceedings of the IEEE International Conference on Data Mining*, November 19–22, Melbourne, Florida, pp. 387–394 (2003).