On Estimation of Associations by using Frailty Distribution

G. G. Shah[1], S. R. Patel[2], N. K. Modi[3]

[1] Faculty of Business Administration, Dharmsinh Desai University, Nadiad, India

[2] Department of Statistics, Sardar Patel University, Vallabh Vidhyanagar, India

[3] Department of Computer Science, Dr. Babasaheb Ambedker Open University, Ahmedabad, India

Abstract: ***In this paper we have considered some associations and their distributions like Multinomial Distributions. We have generated the simulated data and considering base line distribution as Multinomial Distributiion and considering frailty variable having Uniform Distribution, we obtained estimates of the parameters of multinomial distribution and frailty variable considering as random effects.***

Keywords: Associations, Frailty, transactions, correlations, Monte Carlo method

1. Introduction

Data mining is a process of extraction of useful information and patterns from huge data. The process of discovering useful knowledge from a huge data is called as Knowledge Discovery in Database(KDD) and which is often referred to as Data Mining. Data mining is a logical process that is used to search through large amount of data in order to find useful data. Data mining is in fact a broad area which combines research in statistics, database, market basket analysis etc.

Association Rules of Mining introduced by R. Agrawal[1] is an important research topic among the various data mining problems. In Knowledge Discovery, Association Rule Mining plays a vital role. Association rules are one of the most important knowledge of data mining's result which can be defined as the relation between the itemsets by given support and confidence in database. The rule of bread and milk maybe is a common behavior of customer's needs and does not more interesting than bread and diapers or diapers and beer[2]. Amongst many researchers who put forward number of metrics of interest for associations, He, Z at all suggested the frame work based on correlation and the framework was to analyse the residue. Yi et al. verified the rules of correlation of association and Q. Liu et al.[3] extended. For frailty distribution used here have been refered from Parekh et al.[6],[7],[8].

In section 2 of this paper we have studied different correlation of associations like partial and multiple correlations and compared them with simple correlation of association of parameters of multinomial distribution and we obtained the different correlations between these associations, the different partial correlation coefficients and different multiple correlations of interest have been calculated and interpreted in Section-2.Taking the underlined distribution of association as multinomial distributions we have generated the data for different sample sizes and obtained the maximum likelihood estimates, (m.l.e.) of the parameters by using frailty distribution as Uniform distribution in section-3 and section-4 is devoted for estimation of parameters of multinomial distribution by least square theory and obtained least square estimates (l.s.e.) and it is compared with m.l.e.

2. Correlation Matrix of different Associations

We consider here some associations of three items (such as computer, virus scanner and printer).The formation of association is done by selecting one primary item and then some combination of other items in meaningful order are done with some meaningful order usage e.g. if computer is main primary item then one of the associations will be computer, printer and virus scanner. Here considering these item as computer (main item),virus scanner and printer. The combination is called Transaction identification (Tid).We want to study the correlation patterns for the following Tid.
Here correlation matrix is found by using the following Table 2.1 transaction database.

Table 2.1

Transaction id	*Item sets*
1	Computer, Virus Scanner
2	Virus scanner
3	-
4	Computer, Virus Scanner, printer
5	Computer

Considering scan database and generating initial matrix, it gives the following table in which each row corresponding to one transaction and column corresponding to items respectively, containing 1 if item present in the corresponding transaction and 0 if the item is not present in the transaction. Following Table 2.2. display the Transaction Database in Binary Form.

Table 2.2

Tid	Computer	Virus Scanner	Printer
1	1	1	0
2	0	1	0
3	0	0	0
4	1	1	1
5	1	0	0
Item Support	0.6	0.6	0.2

From the above table, we can derive the associations and obtain probabilities. Following Table 2.3 display the Associations Probability

Table 2.3

No	Association	Probability
1	Computer →Virus Scanner	0.4
2	Computer→ Printer	0.2
3	Computer→(Virus Scanner, Printer)	0.2
4	(Printer, Virus Scanner)→Computer	0.2

In the usual notations for the above association pattern support,confidence and lift of the associations have been obtained as under.

Support(Computer →VirusScanner)

$$= \frac{Number\ \ of\ Transactions\ \ of\ Computer\ \ and\ Virus\ Scanner}{Total\ \ Number\ \ of\ Transactions}$$

= 2/5

= 0.4 (2.1)

Using the result of table 2.2 for item support and using the correlation coefficient formula used by Xinog et al.[4] as under for different associations.

Correlation Coefficient between Computer and Virus Scanner defined as

r=

$$\frac{Sup\,(Computer\ ,Virus\ Scanner\ ) - Sup\,(Computer\ )x\ Sup\,(Virus\ Scanner\ )}{\sqrt{Sup\,(Computer\ )x\ Sup\,(Virus\ Scanner\ )\big(1 - Sup\,(Compu\ ter\,)\big)x\,(1 - Sup\,(Virus\ Scanner\ ))}}$$

(2.2)

Using (2.2) and support given in table 2.1 we get correlation between Computer and Virus Scanner r_{cv} association as

$r_{cv} = \frac{0.4 - 0.6\ x\ 0.6}{\sqrt{0.6\ x\ 0.6}\,\sqrt{0.4\ x\ 0.4}}$, in virtue of values from table 2.2 and table 2.3

= 0.167 (2.3)

Similarly with the use of table values from table 2.2 and table 2.3 calculation of other correlation coefficient between Computer and Printer r_{cp} is

r_{cp} = 0.408 (2.4)

and correlation coefficient between Virus Scanner and Printer r_{vp} is

r_{vp} = 0.408 (2.5)

Thus we represent (2.3),(2.4) and (2.5) in tabular form in the following Table 2.4 and display the Correlation matrix of Computer, Virus Scanner and Printer associations as

Table 2.4

	Computer	Virus Scanner	Printer
Computer	1		
Virus Scanner	0.167	1	
Printer	0.408	0.408	1

Denoting associations A for (Computer→Virus Scanner),B for (Computer→Printer),C for (Computer→Virus Scanner, Printer) and D for (Virus Scanner, Printer)→Computer, let event A occur n_1 times with probability p_1,B occur n_2 times with probability p_2,C occur n_3 times with probability p_3 and D occur n_4 times with probability p_4, and let n be total number of individuals selecting different associations. This follows multinomial distribution (n,p_1,p_2,p_3,p_4) with probability mass function.

$$\frac{n!}{n_{1!}\,n_{2!}\ n_{3!}\,n_{4!}}\ {p_1}^{n_1}\,{p_2}^{n_2}\,{p_3}^{n_3}\ {p_4}^{n_4} \quad \text{where } \textstyle\sum_{i=1}^{4} p_i = 1,$$

$p_i \geq 0, \sum_{i=1}^{4} n_i = n$ (2,6)

Partial and Multiple correlation coefficient of different associations: By denoting A by 1,B by 2,C by 3 and D by 4, the partial correlation coefficient between x_i and x_j when x_{q+1} … …,x_p are held fixed $r_{ij.k}$,i,j=1,2, …r, k=q+1 … p and multiple correlation coefficient between X_i and X_{q+1}, …,X_k is ζ_i (q+1,…k),i=2,3…r, j=r+1…k Kshirsagar([5],P.21) had been obtained for multinomial distribution (n,p_1,p_2,p_3,p_4) by using the results of

$$r_{12.3} = \frac{\sqrt{p_1 p_2}}{\sqrt{(1-p_1-p_3)(1-p_2-p_3)}} \tag{2.7}$$

$$r_{13.2} = \frac{\sqrt{p_1 p_3}}{\sqrt{1-p_1-p_3}\,\sqrt{1-p_3-p_2}} \tag{2.8}$$

$$\zeta^2_{1(23)} = \frac{p_1\,(p_2+p_3\,)}{(1-p_1)(1-p_{2-}\,p_3)} \tag{2.9}$$

$$\zeta^2_{2(13)} = \frac{p_2\,(\,p_1+p_3\,)}{(1-p_2)(1-p_1-p_3)} \tag{2.10}$$

$$\zeta^2_{3(12)} = \frac{p_3\,(p_1+p_2)}{(1-p_3)\,(1-p_1-p_2)} \tag{2.11}$$

Remarks : We note that partial and multiple correlation coefficient of multinomial distribution are independent of number of trials.

Thus the partial correlation coefficients for multinomial distribution (n,p_1,p_2,p_3,p_4) are now

$r_{12.3}$=0.57735 ,using (2.7) (2.12)

$r_{13.2}$=0.57735 ,using (2.8) (2.13)

and multiple correlation coefficients are

$\zeta^2_{1(23)}$ = 0.4445,using (2.9) , $\zeta_{1(23)}$ = 0.667

$\zeta^2_{2(13)}$=0.375, using (2.10) , $\zeta_{2(13)}$ = 0.613

$\zeta^2_{3(12)}$=0.375, using (2.11), $\zeta_{3(12)}$=0.613

Comparison of (2.12) with (2.3) shows that the correction coefficient reduces when the third association is ignored, that is partial correlation coefficient is less than the total correlation coefficient. Further the total effect of second and third associations on the first association in 66.7% which is low and hence some other influencing variable should be considered.

3. Estimation by Simulation

In section 2 we have defined multinomial distribution in (2.6) for the different events (different associations).But as the distribution (2.6) is singular, we may not use for estimation purpose.So we define non-singular multinomial distribution such as

$P(N=n_1,N=n_2,N=n_3,p_1,p_2,p_3)$ =

$$\frac{n!}{n_{1!}n_{2!}n_{3!}n_{4!}}\,{p_1}^{n_1}\,{p_2}^{n_2}\,{p_3}^{n_3}\,(1-p_1-p_2-p_3\,)^{\,n-n_1-n_2-n_3}$$

where $p_1+p_2+p_3 \leq 1$, $n_1+n_2+n_3 \leq n$, $0 \leq p_i \leq 1$, i=1,2,3

The association probabilities for different associations obtained in table 2.3 are related to each other like $p_1=2p_2=2p_3$ and $p_4=1-p_1-p_2-p_3$.Let us generate 7000 random numbers, taking 0≤r≤1 as arbitrary independent numbers which can be obtained by using R-Language

With the use of
$p_1+p_2+p_3 \leq r$
$p_2=p_3$
$p_1=2p_2, p_4=1-p_1-p_2-p_3$
generate different sets of (p_1,p_2,p_3,p_4) and using Monte-Carlo (M.C) method and obtaining $(\overline{p_1},\overline{p_2},\overline{p_3},\overline{p_4})$ maximum likelihood estimate of (p_1,p_2,p_3,p_4)

$(\overline{p_i}= \frac{\sum_{ij=1}^{n} p_{ij}}{n}$, i=1,2,3) .

This is done by writing likelihood L as

$$L= \frac{n!}{n_{1}!n_{2}!n_{3}!n_{4}!} {p_1}^{n_1} {p_2}^{n_2} {p_3}^{n_3} (1-p_1-p_2-p_3)^{n-n_1-n_2-n_3} \quad (3.1)$$

and using $\widehat{p_i} = \overline{p_i}$ i=1,2,3 as the solution of maximum likelihood equations obtained from (3.1) ,one can get maximum likelihood estimates (m.l.e) $(\widehat{n_1},\widehat{n_2},\widehat{n_3},\widehat{n_4})$ of (n_1,n_2,n_3,n_4)

Illustration : Let n=7000
Using R Language and M.C. estimate of (p_1,p_2,p_3,p_4) is $(\overline{p_1},\overline{p_2},\overline{p_3},\overline{p_4})$ and for n=7000, $(\overline{p_1},\overline{p_2},\overline{p_3},\overline{p_4})$ is (0.2760,0.1380,0.1380,0.4480) and by solving the maximum likelihood equations and substituting the value of
$(\widehat{p_1} , \widehat{p_2} , \widehat{p_3} , \widehat{p_4})$ and n=7000, we get m.l.e. of n_1,n_2,n_3,n_4 as $\widehat{n_1}$=1932, $\widehat{n_2} = 966, \widehat{n_3} = 966, \widehat{n_4} = 3136$

Similarly by using R-Language and M.C. method we give the following Tables 3.1 and 3.2 showing estimates of (p_1,p_2,p_3,p_4) and (n_1,n_2,n_3,n_4) as $(\widehat{p_1},\widehat{p_2} , \widehat{p_3} , \widehat{p_4})$ and $(\widehat{n_1}, \widehat{n_2} , \widehat{n_3} , \widehat{n_4}$) respectively,

Table 3.1 showing estimate $(\widehat{p_1} , \widehat{p_2} , \widehat{p_3} , \widehat{p_4})$ of (p_1,p_2,p_3,p_4) after generating them

Table 3.1

N	$\widehat{P_1}$	$\widehat{P_2}$	$\widehat{P_3}$	$\widehat{P_4}$
50	0.2915	0.1458	0.1458	0.4169
100	0.2782	0.1391	0.1391	0.4436
200	0.274	0.137	0.137	0.4521
300	0.2683	0.1342	0.1342	0.4633
400	0.282	0.141	0.141	0.436
500	0.2736	0.1368	0.1368	0.4529
1000	0.2276	0.1138	0.1138	0.5449
2000	0.2708	0.1254	0.1254	0.4985
4000	0.2517	0.1258	0.1258	0.4966
5000	0.2493	0.1246	0.1246	0.5014
6000	0.2622	0.1311	0.1311	0.4756
7000	0.276	0.138	0.138	0.448
8000	0.2495	0.1248	0.1248	0.5009
9000	0.2483	0.1242	0.1242	0.5043
10000	0.2276	0.1138	0.1138	0.5448

Table 3.2 showing m.l.e $(\widehat{n_1},\widehat{n_2},\widehat{n_3},\widehat{n_4})$ of (n_1,n_2,n_3,n_4)

Table 3.2

$\widehat{n_1}$	$\widehat{n_2}$	$\widehat{n_3}$	$\widehat{n_4}$	n
15	7	7	21	50
28	14	14	44	100
55	27	27	91	200
81	40	40	139	300
113	56	56	175	400
137	68	68	227	500
228	114	114	544	1000
543	250	250	957	2000
1009	503	503	1985	4000
1243	625	625	2506	5000
1532	787	787	2894	6000
1932	966	966	3136	7000
2001	999	999	4001	8000
2232	1116	1116	4536	9000
2281	1138	1138	5443	10000

By taking different sample sizes the simulated associations (n,n_1,n_2,n_3,n_4) have been obtained in above table No. 3.2

Real Data

We contacted dealer of HP Computer in Nadiad city and they provided us real data of sailing computers for the year 2012-13 as under.

No. of main Computers -	288
No. of associated Virus Scanners	143
No. of associated Printers	141
Others accessories	049

Total	621

Taking aribitraily random number r (0<r<1) the simulated associations given the vector of computer parts as

No. of main Computers	285
No. of Virus Scanners	142
No. of Printers	142
Others accessories	052

Thus for the real data of associations of (Computer,Virus Scanner,Printer) has been obtained by using fraity distribution and they have been compared with real association in the following table.

Table 3.3 showing real data of 621 computers association and frailty simulated values.

Table 3.3

Associations	Real Associations	Simulated Data
Computer	288	285
Virus Scanner	143	142
Printer	141	142
Others	049	052

Above Table shows that the real association and simulated frailty association are very close.

4. Estimation by least square theory

Again using simulated observations of sample size 500 and repeated for 15 times, we have following data showing simulated associations (n_1,n_2,n_3,n_4). Table 4.1 showing simulated n_1,n_2,n_3 of 500 sample size of fifteen times

Table 4.1

n_1	n_2	n_3	n_4	N
150	70	70	210	500
140	70	70	220	500
138	68	68	228	500
135	67	67	232	500
141	70	70	219	500
137	68	68	227	500
114	57	57	272	500
136	63	63	239	500
126	63	63	248	500
124	63	63	251	500
128	66	66	241	500
138	69	69	224	500
125	62	62	250	500
124	62	62	252	500
114	57	57	272	500

By using least square theory for dependent variable n_1 of the associate (n_1,n_2,n_3) as a linear function of associate items (n_2,n_3) as
$n_1=\alpha+\beta_2 n_2+\beta_3 n_3+\in$
where (α,β_2,β_3) are parameters and $\in$ as error term.

By minimizing
$\sum (n_1- \alpha-\beta_2 n_2-\beta_3 n_3)^2$
and solving the normal equations we get least square estimates $(\tilde{\alpha},\widehat{\beta_2},\widehat{\beta_3})$ of (α,β_2,β_3) as under
$\hat{\alpha}$ = -8.221761
$\widehat{\beta_2}$=1.075638
$\widehat{\beta_3}$=1.075638

So that the least square estimates of associations (n_1,n_2,n_3) has been obtained as
$\widetilde{n_1}$ = -8.221761+1.075638 $\widetilde{n_2}$ + 1.07563$\widetilde{n_3}$
where $\widetilde{n_1}$, $\widetilde{n_2}$, $\widetilde{n_3}$ are least square estimates of associations (n_1,n_2,n_3)

Using table 4.1, we get the estimated values of n_1 as n_1, as follows in the vector form
$\widetilde{n_1}$ = (142, 142,137,135,142,138,114,126,127,126,133,140,126,125,114)’

Following table 4.2 showing least square estimates n_1 of n_1 and we obtained x^2 value

Table 4.2

n_1	$\widetilde{n_1}$
150	142
140	142
138	137
135	135
141	142
137	138
114	114
136	126
126	127
124	126
128	133
138	140
125	126
124	125
114	114

x^2_{cal} = 1.4807 < 6.57 = x^2_{14} (0.05)

which show that the simulated values of n_1 in the associations (n_1,n_2,n_3,n_4) are close to estimated values of n_1. Table 4.3 showing comparison of m.l.e. and l.s.e.

Table 4.3

Variable	m.l.e
n_1	137
n_2	68
n_3	68

We note that the m.l.e. and l.s.e. for simulated frailty distribution are close to each other.

5. Conclusions

We obtained associations correlations of interest for different transactions of associations and studied some associations for their frailty distributions and obtained m.l.e. as well as l.s.e. remaining associations may be studied in the similar manner.

References

[1] Agrawal,R.,Imielinski,T. And Swami,A., Mining Association Rules between Sets of Items in Large Databases. In Proc. Of the ACM SIGMOD Conference on Management of Data,Washington D.C., May 1993.

[2] R.B. Fajriya Hakim, New Method to Mining Association Rules using Multi-Layer Matrix Quadrant,

[3] Yi, W.G., Lu, M.Y., Liu, Z.: Regression Analysis of the Number of Association Rules. International Journal of Automation and Computing 8, 78–82 (2011)

[4] Xinog Hui,Shekhar Shahshi,TAPER:A Two-Step Approach for All-strong-pairs Correlation Query in Large Database. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL., X. NO X., XXX 200X

[5] Kshirsagar.A.M. Multivariate Analysis Marcel Dekker, INC 1972.

[6] Parekh, S.G, Patel S.R, Ghosh, D.K,Raykundaliya D.P Discrete Frailty models. Proceddings of the 10th Indiacom : INDIACOM-2016 IEE conference ID:37465 2016 3rd International conference on “Computing for Subs,tainable Global Development”,180-182 (2016)

[7] Parekh S.G,Ghosh D.K., Patel S.R., Some Baysian Frailty models, Inernational Journal of Science and Research

[8] Parekh S.G,Ghosh D.K., Patel S.R., On frailty models for Kidney infection data with exponential baseline distribution. Inernational Journal of Applied Mathematical and Statistics Sciences : 4(s), 31-40, (2005).