STAT30270 – STATISTICAL MACHINE LEARNING (OL)

STUDENT NO - 18201501

ASSIGNMENT - 1

Introduction:

The National Long Term Care Survey (NLTCS) data with record of 21574 elderly people in the United States of America each having disabilities in 16 tasks of daily living has been selected to perform association rule analysis. Also the data holds 3152 rows and 18 columns. Additionally, the recorded disabilities are represented with dummy variables as below.

Variable	Description	Variable	Description
Y1	Eating	Y9	Doing laundry
Y2	getting in/out of bed	Y10	Cooking
Y3	Getting around inside	Y11	Grocery shopping
Y4	Dressing	Y12	Getting about outside
Y5	Bathing	Y13	Travelling
Y6	Getting to bedroom/ using toilet	Y14	Managing money
Y7	Doing heavy house works	Y15	Taking medicine
Y8	Doing light house works	Y16	Telephoning

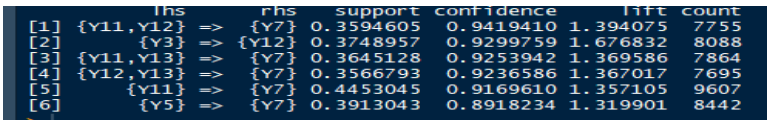
Fig (i): Disabilities in the NLTCS data

Abstract:

Association rules is the most preferred rule to study the dependence between the items selections to determine which are selected together frequently. Moreover to identify the frequent item sets of association rules, we prefer working with Apriori algorithm. Apriori algorithm assists us in minimizing the number of rules from the total possible rules of $3^{items} - 2^{items+1} + 1$. Additionally, Apriori algorithm makes use of interest measures "support", "confidence" and "lift" as threshold to shortlist the most associated frequent item sets.

Initially, embarked the "arules" class data as input to the Apriori algorithm without parameter list for having a clearance in selecting the threshold value of confidence (0.8) and support (0.1). Later, continued with the selected threshold value (confidence = 0.8, support=0.1) that returned us with 17,297 rules. Finally, as we ended up with more rules, we proceed with increasing the threshold parameter support to 0.35 that returns up with reasonable minimum rules of 15 with minimum length of 1 and maximum length of 10 after which the Apriori algorithm stops and returns 0 rules. The reason for specifying the minimum length to be greater than 1 is because the rules with only one item (an empty LHS/ antecedent) will be created, for which both the both the value of confidence and support will be same. But, with our transaction modifying the minimum length to 2 does not reflect any changes to rules because it does not hold an empty LHS rule.

From the 15 rules obtained, we focus on removing the redundant rule, which is nothing but a rule that is observed as a general rule with same or higher confidence values. There is no redundant rules found in the set of 15 rules shortlist using the above threshold values. In case, if redundancy is found, the rule with higher confidence is retained and the others are removed. For instance, the rule with lhs {Y3, Y7, Y11} given rhs {Y5} and lhs rule {Y3, Y7, Y11, Y12} given rhs {Y5} are said to be redundant. It literally means, the elderly people who have disabilities in getting around inside (Y3), lifting heavy weights (Y7) and shopping groceries (Y11) tends to face disability in bathing (Y5) , this rule is obviously, the subset of other rule that deals with elderly people who show disabilities in getting around inside (Y3), lifting heavy weights (Y7), shopping groceries (Y11) and getting about outside (Y12) have disability in bathing (Y5). Additionally, if the confidence of the first rule (0.8255) is greater than confidence of second rule (0.8233), first rule is retained whereas the second rule is removed.



	lhs		rhs	support	confidence	lift	count
[1]	{Y11,Y12}	=>	{Y7}	0.3594605	0.9419410	1.394075	7755
[2]	{Y3}	=>	{Y12}	0.3748957	0.9299759	1.676832	8088
[3]	{Y11,Y13}	=>	{Y7}	0.3645128	0.9253942	1.369586	7864
[4]	{Y12,Y13}	=>	{Y7}	0.3566793	0.9236586	1.367017	7695
[5]	{Y11}	=>	{Y7}	0.4453045	0.9169610	1.357105	9607
[6]	{Y5}	=>	{Y7}	0.3913043	0.8918234	1.319901	8442

Figure (ii).Top five rule sorted with confidence measure

As a result of the Apriori algorithm, the top 5 rules out of 15 association rules were sorted with respect to confidence measure as listed in the figure (ii) above (Confidence - A measure of probability of consequent (RHS) given antecedent (LHS)). From the figure (ii), the rule #1 reveals that 7755 elderly people have 94% probability of having disabilities in doing heavy house works (Y7) given they have disability in doing grocery shopping (Y11)and getting about outside (Y12). In this case, interest measure "support" explains that the combination of lhs and rhs in rule #1 co-occurs with the probability percentage of 35%.

Additionally, "Lift" measure 1.394 indicates about the strength of association rule. Since, the lift > 1, it states that association rules holds good for rule #1. Similarly, rule #2 explains that 8088 elderly people who show difficulties in getting around inside (Y3) will have difficulties in getting about outside (Y12) with a confidence value of 93% and with stronger association rule represented using its lift value 1.68 and 37% probability in both combination occurring together. In rule #3, elderly people have disabilities in doing heavy house works (Y7) given if they express disability in grocery shopping (Y11) and traveling tends (Y13) was accounted by 7864 people. Also, there is 36% probability that these antecedent and consequent have probability of occurring together. Moreover, Rule #4 explains that there are 7695 elderly people having difficulty in getting around outside (Y12) and traveling (Y13) have difficulty in doing heavy house work (Y7) with a 92% confidence and with 35% dependence between item sets lhs and rhs. Finally, from rule #5, we are 91% confident that elderly people, expressing disability in doing heavy house works (Y7) given they express disabilities in grocery shopping (Y11) and the probability of their co-occurrence is 44% and indicates good strength of association rule between the rule #5 lhs and rhs with 1.35 value of lift. The below plot Fig.2 explains the relationship between support and confidence measure with respect to lift based on the 15 rules mined. Fig.2 depicts that as lift increases, support decreases and confidence increases.

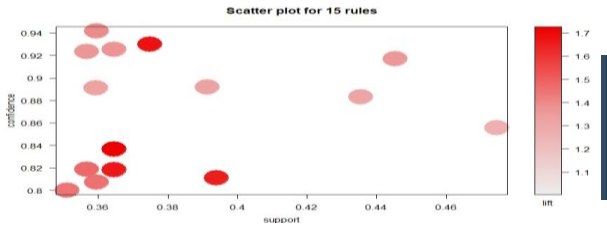


Fig2. Top 5 rules sorted on standardized lift

	lhs	rhs	support	confidence	lift	count	PA	PB	LB	UB	sLift
[1]	{Y11,Y12}	=> {Y7}	0.3594605	0.9419410	1.394075	7755	0.3816168	0.6756744	0.2221892	1.480003	0.9316850
[2]	{Y3}	=> {Y12}	0.3748957	0.9299759	1.676832	8088	0.4031241	0.5546028	0.0000000	1.803092	0.9299759
[3]	{Y11,Y13}	=> {Y7}	0.3645128	0.9253942	1.369586	7864	0.3939001	0.6756744	0.2614126	1.480003	0.9093897
[4]	{Y12,Y13}	=> {Y7}	0.3566793	0.9236586	1.367017	7695	0.3861593	0.6756744	0.2369852	1.480003	0.9091039
[5]	{Y11}	=> {Y7}	0.4453045	0.9169610	1.357105	9607	0.4856309	0.6756744	0.4915920	1.480003	0.8756610
[6]	{Y3}	=> {Y7}	0.3592751	0.8912269	1.319018	7751	0.4031241	0.6756744	0.2892957	1.480003	0.8647992

Fig3. Top 5 rules sorted based on the standardized lift

The lift of each rule and their respective value of lower and upper bound are calculated using Frechet rule is updated in the below table Fig (iii) as LB and UB respectively. Hence, all the values of lift in the below adhere and falls within that range, where the lower bound is 0.222 and the upper bound is 1.480 for the top rule sorted by standardized lift. Sorting the rules with standardized lift offers a natural and unambiguous method of ranking the association rules. Also, standardized lift makes the bound tighter. From the table, it is noted that we have the same set of top 5 rule even after sorting the rules with respect to standardized lift.

Inference:

- On experimenting with different threshold values, the best selection of interest measure is found to be with confidence value equal to 0.8 and support value equal to 0.2 by trying to modify the rules varying the value of interest measure threshold confidence and support values.
- The top rule by sorting with confidence revels that elderly people tend to have disability with doing heavy house work given they have disability in grocery shopping (Y11) and getting about outside (Y12).
- When considered with support, the top rule explains that there is 44% probability of elderly people with disability in grocery shopping (Y11) and tends to have disability in doing heavy house work (Y7).
- Moreover, top rule on sorting with lift value reveals that elderly people with disability getting around inside (Y3) has good strength of association with disability in getting about outside.
- Since, all lift values and standardized values updated above 1 and nearing to the value 1, the association rules are reported to be stronger associate rule and indicative of interesting rules respectively. To reiterate, the association rules obtained by the above threshold values are found to be of good choice from its lift and standardized lift values.