Data Mining In Conceptualising Active Ageing

Richi Nayak[1], Laurie Buys[2] and Jan Lovie-Kitchin[3]

[1]School of Information Systems, [2] Centre for Social Change Research, [3] Faculty of Health
Queensland University of Technology, Brisbane, Australia

{r.nayak, l.buys, j.kitchin@qut.edu.au}

Abstract

The concept of older adults contributing to society in a meaningful way has been termed 'active ageing'. We present applications of data mining techniques on the active ageing data collected via a survey of older australian on a wide range of social and behavioural variables. The goal is to understand the underlying relationships and attributes which characterise active ageing. The data mining results indicate that an individual's health, attitude to learning, social network support and (positive) emotional feelings are significant contributors to achieving active ageing.

Keywords: Classification, clustering, association mining, Active ageing

1 Introduction

The concept of older adults contributing to society in a meaningful way has been termed 'active ageing' (Kinsella and Phillips 2005). The policy framework for active ageing developed by the WHO emphasizes that health, participation and security are important for quality of life for older adults (WHO 2005). This framework is guiding the conceptual development of "active ageing" in diverse national contexts.

To conceptualise active ageing in terms of complex issues that intertwine and converge with the ageing experience, rather than in a singular health/social dimension, the Australian Active Ageing (Triple A) study at the Queensland University of Technology (QUT) has conducted a national-wide postal survey to collect the responses of older people on a wide range of questions related to 'work', 'learning', 'social', 'spiritual', 'emotional', 'health and home', 'life events' and 'demographics'. Previously, studies such as the Australian Longitudinal Study of Ageing (ALSA) (Andrews, Clark, and Luszcz 2002), and the Dubbo Study of Ageing (Simons et al. 1990) included social aspects as well as the more usual psychological and behavioral (ALSA), and bio-medical issues (Dubbo Study). This study at QUT is the first of its kind reflecting a wide variety of aspects of older adults life.

Data Mining (DM) techniques have been successfully applied to a number of application domains including finance, marketing, health, Internet and others (Han 2001). This paper extends the use of DM to understand how the older population in Australia is ageing. The large number of interrelated variables and the complex relationships between various life aspects of older people's experience greatly enhance the potential of applying data mining (DM) techniques.

We use different data mining tools available within the 'SAS 9.1 Enterprise Miner' to conduct experiments. In particular, we used clustering, predictive modelling and association mining to carry out an exploratory analysis of the data. The results successfully highlight interesting trends in the data and describe that an individual's health, attitude to learning, social network support and (positive) emotional feelings are significant contributors to achieving active ageing. The trends and patterns of this data set will assist to develop the concept of active ageing and contribute significantly to future policy directions.

2 Data Mining Application to Active Ageing

The goal is to pre-process the national survey data; apply an appropriate DM technique or a combination of techniques; and evaluate and interpret the meaningful rules to enhance the understanding of active ageing.

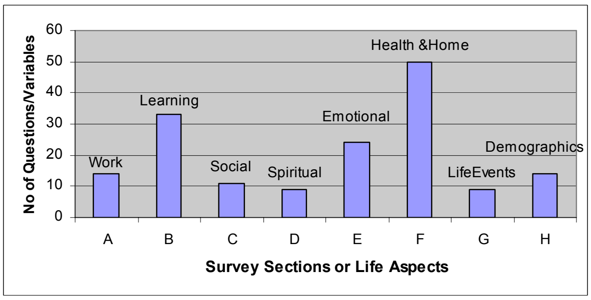

Figure 1: Distribution of Variables

2.1. Data Pre-processing

The first task was to understand the nature of the data set to decide the strategies and methods of data mining. The Triple A study data includes responses of members of a large Australia wide seniors organisation on a wide range of questions related to their life. A total of 2655 surveys were returned at 46% response rate. Of these 32 incomplete surveys were excluded due to missing age, postcode and/or non-response to survey questions; others were excluded due to respondents being younger than 50 years old.

Survey questions were divided into groups/aspects such as 'work', learning', 'social', 'spiritual', 'emotional', 'health', 'home', 'life events' and 'demographics'. Each question is treated as a variable/attribute and its responses become the values of the variable in the data mining analysis. Figure 1 lists the number of variables in each aspect. There were many transformations done in the data set in order to analyse with data mining algorithms.

2.1.1 Empty Fields

A major problem in preparing data for mining was that some questions requested the respondents to skip questions depending on their response to certain questions. These variables were subsequently modified to limit the effect of the large percentage of respondents who didn't answer them. For example, variable A1 requested respondents to skip to A4 if they answered no, which led to a large percentage of respondents leaving variables A2 and A3 blank. Therefore A2 and A3 were modified to include a response of N/A for those respondents who were not required to do so.

2.1.2 Value Transformation

There were some variables that were transformed to fulfil the goal of the mining process. For example, the variable H2 asked respondents to supply their age. Since the Triple A project is only interested in analysing the various age groups, therefore the data was modified to reflect the group that the age variable belongs to (55-64; 65-74; 75 years and over). Similarly, the variable H3 asked respondents to supply their postcode. We found that the postcode had little effect on the data, and therefore used it to calculate the state that the respondent lived in, whether they lived in a metropolitan or regional area, and to which socio-economic status they belong.

Finally, the processed data consists of 2,623 cases and 165 variables. Majority of variables are categorical. The distribution analysis showed that the data set is suitable for any data mining technique to be applied. The result of the data mining would highly depend upon how we formulate the problem.

2.2 Conducting Data Mining

The main objective in this study was to find the inter-relation between various aspects of life to achieve active ageing. We decided to first apply the clustering analysis to understand the nature of the data set overall. We utilized the most commonly used centroid based approach *k-means* (Jain, Murty and Flynn 1999). Next, we applied the association and predictive modelling to understand the deeper relationships between various sections and various variables.

We had many choices for the predictive modelling (Lim and Loh 2000). Since, our main goal is to understand how the variables from different life aspects are related to each other, good comprehensibility and high accuracy were the main requirements in a method. *Decision tree* is chosen due to their ability to obtain reasonable accuracy, good comprehensibility, efficiency, robustness and scalability.

The *Apriori algorithm* is used for the association analysis. As support is lowered the processing time to find association rules dramatically increases and it is often infeasible to find all association rules. Consequently only maximal association rules are chosen as it reduces the number of rules that need to be analyzed. A maximal association rule is a supported association rule that isn't contained in another supported association rule (Agrawal, Imielinski and Swami 1994). Typically interesting maximal association rules are only supported by a low percentage of entries.

3. Analysis of generated clusters

To understand if there is a clear segmentation in the data set, clustering analysis is performed. The k-means clustering algorithm divided the data set into 7 distinct but overlapped clusters. There is no segmentation of the value ranges of the variables, such that all values within a continuous interval of each variable belong to the same cluster. The clusters are quite close to each other. This indicates that variables don't define distinct boundaries, though the given variables are slightly more interesting than others. There exist some highly correlated relationships within the data which should lead to interesting rules for active ageing. The figure 2 shows the number of respondents grouped in 7 clusters.

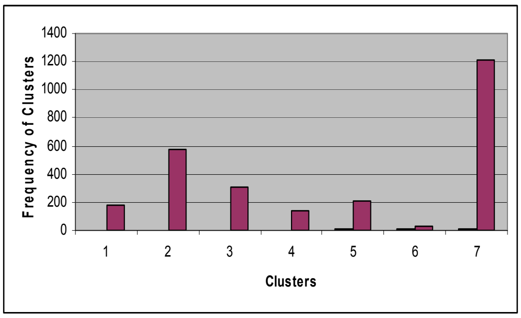

Figure 2: Cluster Distribution

The cluster 7 that groups the majority of respondents showed that the people feel contented in the current life when they enjoy the social interactions, and they have more intention to learning new things. Also, health did not limit them in doing the outside activities.

The cluster 2 grouped the people who do not feel contented in their current life. Respondents in this group have lower social interaction, and are less capable of handling their social relationships .The cluster 3 grouped the people according to the difficulty level at performing daily activity. Majority of people were from the age cohort 75 years and over, and some from 65-74 years.

The figure 3 shows (1) the distribution of the 44 variables that appear in clusters across each aspect of life, and (2) the distribution of the 44 cluster-variables in each aspect in comparison to the original distribution of the total 165 variables in each aspect of the data set. For example, the aspect B contributes 27% (12 out of 44) of variables in the total number of variables appearing in the clusters, whereas the entire data set includes only 33 (i.e. 20%) aspect B variables (figure 1).

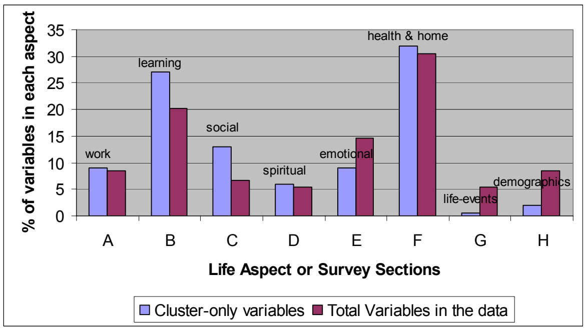

Figure 3: Distribution of Variables in Clusters

The figure 3 shows that the 'Health & Home' and 'Learning' are the most prominent in terms of number of variables in clusters. This indicates that an individual's health and attitude to learning are the most significant contributors to active ageing. However, when comparing the cluster variables and original variable distribution, the figure 3 shows that the variables on 'Social', 'Learning' and 'Health & Home' dimensions have more impact, since their proportion in the clusters is much higher than the original.

The Variable D4, which asked whether the responders feel contented in their present life, is found to be the most influential to form clusters based on the relative performance measure of the variables in determining the clusters.

The results of the cluster analysis indicate that there are correlated patterns within the data. However it is necessary to perform the predictive modelling or association analysis on a selection of variables to determine whether those correlations have any meaning.

	# Variables		# Rules		Misclassification Rate	
Selected target from all targets	All	Cluster	All	Cluster	All	Cluster
A1: Full time or part time paid work	29	33	57	90	3.33%	4.06%
A2: #Hours paid work per week	79	43	145	192	11.79%	19.53%
A12: Voluntary activities or Not	64	38	116	143	9.26%	18.14%
A13: Involved in tertiary degree courses or diplomas	84	42	156	206	13.49%	16.43%
B4: Interested in new or current event program	98	41	196	206	21.30%	27.32%
B10: Extent to open to new technology	79	41	187	222	15.13%	19.55%
B11: Need to learn to enjoy learning new things	78	39	191	190	14.66%	16.70%
B12: Need to learn to organise holiday/travel arrangements	82	39	183	204	23.74%	38.38%
C1: # (Emotionally) close People within one hour travel away	109	43	207	274	55.02%	55.59%
C3: #family and friends who had phone conversation in the past week	113	42	231	245	43.35%	46.91%
C5: Family and friends understands the respondent or Not	103	42	209	217	14.78%	18.81%
C6: Feel useful to family or friends or Not	112	42	225	225	17.28%	20.51%
D1: Believe in a higher being or Not	104	43	207	262	30.50%	40.37%
D3: Feel in control of your life or Not	95	42	208	230	20.15%	23.31%
D4: Feel contented in your life	110	43	226	223	20.31%	24.68%
D9: Feel as searching for personal meaning	108	42	221	246	36.32%	37.81%
E1: Confidence in the opinion or Not	105	42	206	228	21.90%	25.55%
E12: Happy with the personality or Not	103	43	214	244	18.70%	23.20%
E23: Satisfied with accomplished in life	113	42	256	230	26.49%	30.63%
F3f: Limitation of health in bending	101	42	215	241	17.65%	22.06%
F3g: Limitation of health in walking	78	42	138	171	10.09%	14.54%
F4b: Limitation of health in accomplishing the wishes	105	43	210	207	14.82%	19.05%
F4c: Limitation of health in work	78	39	160	166	11.87%	14.23%
G1:Effect of a major illness in life	97	41	205	237	16.78%	20.46%
G2: Effect of change of work in life	90	41	208	209	19.20%	23.19%
G3: Effect of a new study course in life	100	41	182	195	16.88%	19.13%
G6: Effect of child-care activities in life	94	40	171	168	12.67%	15.16%
H1: Gender	92	42	186	225	13.99%	18.77%
H2: Age group	122	42	222	274	17.19%	20.18%
H3a: Residential State in Australia	117	41	228	211	39.74%	45.54%
H4: Country of Birth	93	42	169	203	12.97%	15.33%

Table 1: Predictive Models Details

4. Analysis of Predictive Models

Predictive modelling or classification is performed to establish relationships that exist between various sections (life aspects) and various variables present in them. We wanted to examine how a variable in one aspect is related to other variables. We have chosen four of the target variables in each life aspect based on (1) the equal distribution among their representative class values, and (2) their appearance in clusters.

We build the decision tree models for each target with two kinds of input attributes: (1) all the rest of attributes as dependent labelled as 'All', and (2) only the rest of the attributes that appear in the clusters labelled as 'Cluster'. The results in Table 1 show the details of the models for each target with the average performance on 10-fold CV experiments. We report the number of variables contained in the rules, the number of classification rules, and the misclassification rate for the classifiers.

The relatively small misclassification rates (26 out of 32 have < 30% misclassification) in Table 1 show that the models can be found that can accurately describe some of the data. This indicates that there are strong patterns within the data. The large range of required variables for the 'All' group in Table 1, and the poor misclassification rates in the models using only 'Cluster' variables (A12, B12, D1 and E16 in particular), indicates that the subset of variables from the cluster are insufficient to accurately describe all of the given targets. This also helps to reveal the nature of overlapped clusters. These results confirm that since respondents have very similar backgrounds (such as similar life style as supported by the statistical analysis of responses), the k-means clustering is not able to categorize them in distinct and disjoint classes. Accordingly, clustering variables alone can not be used for building a classifier model.

The Figure 4 illustrates how each life aspect influences in building predictive models by showing the percentage of variables from each aspect that form the model for the given target. The average of the targets for each aspect is calculated by dividing the number of variables for the given aspect by the number of variables required to classify the target. We compare the percentage of variables that exist in each section (or life aspect) in the survey with the mean percentage of dependence for each aspect based on experiments. The Figure 4 shows that the Learning, Emotional, and Health and Home aspects (B, E, F) are the most significant, as combined they contribute 64.49% of all classification variables. This indicates that physical and emotional health combined with the desire to learn are the most significant factors when considering active ageing. However, these aspects also contribute 65.24% of the total variables in the data set so this could be expected. It is therefore necessary to conduct further analysis to measure the significance of each life aspect.

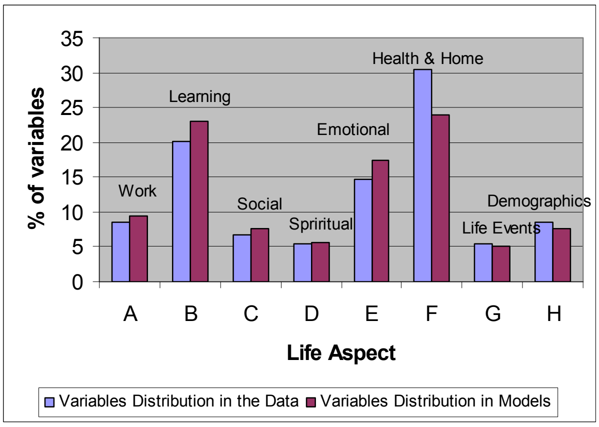

Figure 4: Average Aspect Dependencies

The Table 2 shows the analysis to indicate the aspects on which some targets (life aspects) rely more so that the variables from those aspects can be deemed important in determining those targets. The results in Table 2 that are shaded dark grey highlight results that have a 20% or more difference from the "Variables" row. The "Variables" row shows the percentage of variables that exist in each life aspect in the data. The "Mean" row shows the mean percentage of dependence for each aspect based on experiments. The highlighted results indicate that the target from the given aspect rely on the prescribed aspect more than is statistically expected. Therefore the variables from those aspects are quite important in determining those targets.

Target	**Section Breakdown**							
Section Average	**A (%) Work**	**B Learning**	**C Social**	**D Spiritual**	**E Emotional**	**F Health & Home**	**G Life Events**	**H Demographics**
Variables	8.54	20.12%	6.71%	5.49%	14.63%	30.49%	5.49%	8.54%
A	12.48%	20.76%	9.30%	5.19%	14.07%	18.46%	7.03%	12.71%
B	10.99%	27.36%	6.20%	5.58%	19.78%	19.15%	5.38%	5.56%
C	8.91%	23.83%	7.99%	5.08%	18.33%	25.86%	4.08%	5.93%
D	8.84%	22.93%	8.02%	6.28%	17.58%	24.17%	4.74%	7.44%
E	6.93%	23.11%	7.42%	6.54%	19.48%	24.23%	4.66%	7.64%
F	8.53%	20.87%	7.32%	5.94%	15.06%	29.88%	4.75%	7.65%
G	9.23%	23.67%	8.47%	4.71%	15.11%	26.29%	6.01%	6.50%
H	9.13%	22.25%	6.65%	5.96%	19.55%	24.10%	4.53%	7.84%
Mean	9.38%	23.10%	7.67%	5.66%	17.37%	24.02%	5.15%	7.66%

Table 2: Detailed Analysis of Dependencies of Each Aspect

Targets from aspect A are more reliant on variables in aspects A, C, G, and H and less reliant on aspect F. Targets from aspect B are more reliant on variables in aspect A, B, and E and less reliant on aspects F and H. Targets from aspect C are more reliant on variables in aspect E, and less reliant on aspects G and H. Targets from aspects D, E, and H are more reliant on variables in aspect E, and less reliant on aspect F. Targets from aspect G are more reliant of variables in aspect H.

Therefore it can be concluded that variables in E (Emotional) aspect are more significant, and variables in F (Health & Home) are less significant than any other aspects (according to the difference in values of "Variables" and "Mean" row). Despite this it is still evident that aspects B, E, and F are the most important in this survey.

To further investigate the dependencies of various life aspects and variables, we generated models to predict the selected target considering only the variables from each aspect. We compare the dependencies of each aspect in building the predictive models independently with the model that includes all variables from all aspects. The misclassification rates of the model based on all variables versus models reliant on only variables from the specified aspect are shown in Figure 5.

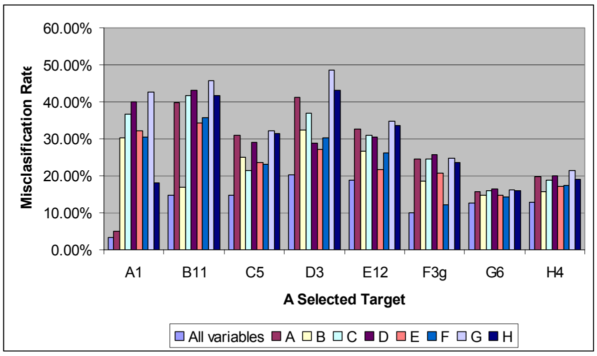

Figure 5: Models versus Life Aspects

The figure 5 shows that every target except G6 and H4 produces significantly more accurate models when only relying on variables from their respective aspect. The G6 and H4 targets can be accurately predicted (<= 21.4% misclassification) using variables from any aspect. A total of 81.81% of respondents answered "No" to G6: "Were they involved in any childcare activities in the past year?" and 76.07% of respondents answered "Australia" to H4: "Country of Birth". Therefore when such a high percentage of respondents answered the same way it is highly likely that strong relationships exist between many variables that can explain the data.

The G6 target is within 30% of accuracy of the base model for every aspect except aspect D, which reiterates the previous conclusion that G6 can be explained by several variables. However, excluding G6, only B11, E12, and F3g have models that rely on their own aspect so strongly. In simpler words, only these models are within 30% of the accuracy of the base classifier. Therefore it is reasonable to assume that strong correlations exist within an aspect. Again 'Learning', 'Emotional', and 'Health and Home' (B, E, and F) appear to be relied upon the most heavily.

5 Analysis of Association Rules

Association rule mining was performed to establish dependencies between various variables in certain selected sections without any preconditions given.

No of Association Rules for the given Support			
	All	Male	Female
100%	0	0	0
90%	44	0	0
80%	636	0	0
70%	5,388	0	0
60%	43,507	0	0
50%	N/A	0	64
40%	N/A	19	3,811
30%	N/A	9,288	N/A

Table 3: Association Rules versus Support

The results in Table 3 show the number of maximal association rules found for the different data sets at the given support requirements. The "All data set" includes all of the respondents whilst the "Male and Female data sets" contain only male and female respondents respectively. The higher number of rules inferred for female than male shows that female respondents are in more agreement. Results for the Female data set with 50% minimum support are supported by 28.3% of respondents, as 56.7% of respondents are female in the survey.

Note that as support is lowered there is a dramatic increase in the number of maximal association rules found. It becomes infeasible after some point due to excessive computation efforts (e.g., at and below 50% support for the All data set). Therefore to find interesting association rules it is necessary to create meaningful subsets such as the Male and Female groups.

Variables from various Aspects (in %)				
	At 90% support	80%	70%	60%
A	0	0	0.2	1.7
B	0	0	4.1	12.3
C	0	0	0	0.3
D	0	0	0	0
E	0	0	0	0.02
F	100	99.8	99.9	99.9
G	25	58.3	72.8	77.5
H	0	6.9	20.9	30

Table 4: Percent of Association Rules Containing Variables from Various Aspects

The results in Table 4 show the percentage of rules that contain at least one variable from the specified section for the All data set. Clearly section F is important as it is contained in over 99% of all maximal association rules for the given tests. Section G is the next most important whist sections A, B, C and H are not as important. However section D has no variable, and section E has only one variable where 60% of respondents gave the same answer. Therefore to find interesting variables containing sections D and E a much lower minimum support value is required. As stated previously though it is infeasible to find all maximal association rules with low support, and therefore specific subsets must be chosen to find interesting association rules within these sections.

6. Discussion and Conclusion

The Triple A project at the Queensland University of Technology, Brisbane, Australia incorporates a wide scope of issues significant to the quality of life for older people. With the use of the data mining techniques, this paper examined the inter-relationships of a wide range of 'work', 'learning', 'social', 'spiritual', 'emotional', 'health and home', 'life events' and 'demographics' variables in order to identify those that contributed most strongly to positive responses and could therefore be indicators of "Active Ageing".

Firstly, clustering was applied to understand the nature of the data set. The respondents were put together into seven groups according to their feeling of 'contented' in their life, physical health limitations, etc. Moreover, the overlapped clusters showed that variables related to the 'Health and Home', 'Learning', 'Social' and 'Emotional' sections are prominent in deciding clusters of common characteristics. This indicates that an individual's health, attitude to learning, social network support and emotional feelings are significant contributors to positive overall wellbeing which is an indicator of active ageing. The results of the cluster analysis indicate the need to perform predictive modelling on a selection of variables to determine whether those correlations have any meaning.

A number of decision tree models were built to further explore the relationships between various variables targeted to certain variables. The results show that the subsets of variables as they appeared in clusters alone are insufficient to accurately describe all of the given targets. This emphasizes that a complex relationship exists between variables to describe a concept, and all the variables do contribute significantly.

We also modelled several decision trees considering all variables to predict a number of targets chosen from each aspect. The analysis of number of variables appearing in various models shows that the 'Learning', 'Emotional', and 'Health and Home' aspects are more significant, as combined they contribute 64.49% of all model variables. Further analysis shows that 'Emotional' feelings are more significant, and 'Health and Home' is less significant than any other aspect. Factors related to 'Learning' are also significant. This was evident when we closely examined the trees. Majority of rules have dominance of 'emotional feelings' and 'learning needs and interests' in explaining active ageing concepts.

Association analysis was also performed to find the rules of conceptualising active ageing without any preconditions or bias given as in predictive modelling. Association rules are inferred based on the rate of agreement in overall responses. The variables related to 'Health and Home' appeared in over 99% of all rules for the given tests. The 'Life Events' information was the next important. This indicates that majority of respondents in this survey do agree on questions about two aspects of life, which is related to the fact that respondents were selected from the same database. The majority shared similar characteristics such as being comfortable financially, living with a companion, in their own home in a metropolitan area and being well educated.

A number of lessons are learned by applying data mining to a sample survey data for the purpose of knowing more about the complex ageing process.

- The survey data includes a homogenous population of respondents in terms of education and finance status. For more insight, we need to mix various populations such as those with diverse living status, lower education, lower financial status etc. Regardless, the data mining techniques were able to capture the essence of the data set to reflect general positive engagement in life. It reinforces the fact that the inferred patterns are as good as the data are.
- Clustering is a good technique to understand the basic nature of the data set. However to understand the data in depth, the application of predictive modelling and association analysis techniques were required.
- The main goal of this data mining application was to understand the survey data and learn if there are any patterns and rules exit that reflect the population. This emphasizes the need for good comprehensibility of output results, which demonstrates the reason that decision tree was chosen for predictive modelling.
- Spending significant time in preprocessing, formulating the various subsets of problem and analyzing the comprehensible results were the main reasons for the success of the project. The data mining output alone is not sufficient; there were a large number of decision tree rules. However, a detailed analysis of these models resulted into a detailed comparison of various life aspects. So the analysis of the output result should get equal importance.
- Data mining analysis successfully highlighted complex issues that intertwine and converge with the ageing experience, rather than in singular health or economic dimensions. Our results present a portrait of Australian older adults that is distinctly different from the stereotype generated by models which negatively portray ageing as a process of decline.

The identified relationships assist us to better understand the impact of ageing upon older Australians' engagement in society and their general sense of well being. The identified patterns will help policy makers and service providers to provide services that meet the needs of older people in the future.

The future research includes (1) the analysis of data with hierarchical clustering as many variables are strongly correlated and (2) the association mining analysis to shift from general criteria to more specific ones by regrouping the variables into interesting types such as communication technologies, TV, car driving etc.

7. Acknowledgement

This work was partly supported by the QUT ECR and the QUT Strategic Collaborative Grant Scheme. We would like to thank all the team members of the Triple A project. We would also like to thank Calum Robertson, Lei Zhang, Xin Li and Ying Luo for their assistance in conducting experiments.

8. References

Agrawal, R., Imielinski, T. and Swami, A. (1993): Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, **22**:207-216, ACM Press.

Andrews G, Clark M, Luszcz M. (2002) Successful aging in the Australian Longitudinal Study of Aging: Applying the MacArthur model cross-nationally. Journal of Social Issues; 58(4): 749-765.

Han J. (2001) *Data Mining: Concepts and Techniques* Morgan Kaufmann Publishers, San Francisco

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys (CSUR), 31*(3), 264-323.

Kinsella K. & Phillips, DR (2005). Global aging: The challenge of success. Population Bulletin 60(1): 3-40.

Lim, T. S. & Loh, W. Y. (2000). A comparison of prediction accuracy, complexity and training time of thirty three old and new classification algorithms. *Machine Learning, 40*(3), sep. 203-228.

Simons LA, McCallum J, Simons J, Powell I, Ruys J, Heller RLC (1990). The Dubbo study: An Australian prospective community study of the health of elderly. Australian & New Zealand Journal of Medicine; 20: 783-789.

World Health Organisation [WHO]. Active Ageing: A Policy Framework. [Online] [Accessed 2005 April]. Available from URL http://www.who.int/hpr/ageing/ActiveAgeingPolicyFrame.pdf