The Use of Various Data Mining and Feature Selection Methods in the Analysis of a Population Survey Dataset

Ellen Pitt and Richi Nayak

Faculty of Information Technology,
Queensland University of Technology
George Street, Brisbane, 4001, Queensland

ea.pitt@student.qut.edu.au and nayak@qut.edu.au

Abstract

This paper reports the results of feature reduction in the analysis of a population based dataset for which there were no specific target variables. All attributes were assessed as potential targets in models derived from the full dataset and from subsets of it. The feature selection methods used were of the filter and wrapper types as well as clustering techniques. The predictive accuracy and the complexity of models based on the reduced datasets for each method were compared both amongst the methods and with those of the complete dataset. Analysis showed a marked similarity in the correlated features chosen by the supervised (filter) methods and moderate consistency in those chosen by the clustering methods (unsupervised). The breadth of distribution of the correlated features amongst the attribute groups was related in large part to the number of attributes selected by the given algorithm or elected by the user. Characteristics related to Health and Home, Paid and Volunteer Work and Demographics were the targets for which predictive accuracy was highest in both the reduced and full datasets. These attributes and a limited number of characteristics from the Learning, Social and Emotional attribute groups were important in clustering the population with Health and Home characteristics being most consistently important. Misclassification rates for models associated with most targets decreased with the use of subsets derived via filter methods but were increased for subsets derived using clustering methods.

Keywords: Data mining, feature reduction, wrapper methods, filter methods, ageing populations, self-assessed health status..

.

1 Introduction

It is known that not all attributes in a multidimensional database are required to solve a given data mining problem and, in fact, the use of some attributes may increase the overall complexity and decrease the efficiency of an algorithm (Dunham, 2003b). Dunham notes that dimensionality can be reduced but a decision regarding which attributes should be retained is difficult. Several authors have considered the best definition for features included in reduced subsets with Dash (1997) defining these as features that are relevant and non-redundant while Kohavi and John (1997) considered the difference between optimal and relevant features. The most commonly used feature selection methods are of two types: wrapper and filter (Tang and Mao, 2005), (Kohavi and John, 1997), (Bengio, 2006)). The filter method selects features based on the general characteristics of the training data while the wrapper methods use a learning algorithm to assess the accuracy of potential subsets in predicting the target. Wrapper methods have higher computational costs (Langley, 1994). Filter methods are often less costly in time and adequate for datasets with large numbers of instances. Filter methods can be further categorised into two groups, namely attribute evaluation algorithms and subset evaluation algorithms, based on whether they rate the relevance of individual features or feature subsets. Attribute evaluation algorithms rank the features individually and assign a weight to each feature according to each feature's degree of relevance to the target feature. Yu and Liu (2003) note that attribute evaluation methods are likely to yield subsets with redundant features since these methods do not measure the correlation between features. Subset evaluation methods, in contrast, select feature subsets and rank them based on certain evaluation criteria and hence are more efficient in removing redundant features. Kohavi and John (1997) note the problems with each method and propose that wrapper methods are better in defining optimal features rather than simply relevant features and that they do this by allowing for the specific biases and heuristics of the learning algorithm and the training set being used.

Feature selection methods involve generation of the subset, evaluation of each subset, criteria for stopping the search and validation procedures (Dash, 1997). The characteristics of the search method used are important with respect to the time efficiency of the feature selection methods. Search methods include BestFirst, Exhaustive, FCBF, Genetic, GreedyStepwise, Race, Random and Ranker. BestFirst is considered by Aha and Bankert (1995) to be superior to GreedyStepwise and better for

subset evaluation while Ranker is more appropriate for attribute evaluation methods. BestFirst differs from the GreedyStepwise in the inclusion of a backtracking component in which the number of non-improving nodes is controlled. Exhaustive searches start with an empty subset whereas GreedyStepwise and BestFirst may start with an empty or full subset or a random subset size (Witten and Frank, 2005).

Clustering allows dataset components to be allocated to particular groups according to the similarity of their features and in the absence of a target variable. Various clustering methods define the importance of each attribute in determining cluster content. The features of importance in determining the content of clusters in the Active Ageing dataset are to be used as a means of feature reduction and the efficiency of such a method is to be compared with the filter and wrapper methods.

Data Mining and feature selection (FS) techniques have been used in medical contexts in a variety of situations, most commonly in gene analysis but also in the study of other clinical, psychosocial and epidemiological issues. For many situations, there is no clearly defined target in the multivariate dataset. Bhargava (1999) discusses a dataset of US Gulf War veterans, with over 20,000 records and 150 variables, and the use of a genetic algorithm to identify subsets that lead to interesting patterns.

The Active Ageing dataset being studied here has been previously analysed using the data mining techniques of clustering and association (Nayak et al., 2006), (Nayak and Buys, 2006) and attributes in the Health and Home, Learning, Social and Emotional groups were considered important in clustering the population. Others have studied ageing populations and have attempted to define characteristics associated with longevity and various quality of life indicators in ageing populations of various countries (Mackenbach et al., 2002) ; (Lyyra et al., 2006); (Lee, 2000), (Kawada, 2003); (Idler et al., 1990); (Maxson et al., 1996); (Nishisaki et al., 1996); (Nybo et al., 2001). Both Nishisaki et al (1996) and Kawada (2003) have shown self reported health status to be an independent predictor of survival.

Svedberg et al. (2005) discussed the changes in self reported health status over a 9 year follow up period in a twin population. They found more substantial cohort differences than longitudinal changes and suggested that "socially mediated and individual-specific environmental effects" had more effect on "phenotypic stability" over the 9 year period than individual differences related to genetically determined diseases. This situation suggested to them that there is opportunity for interventions that would maintain a population's self reported health status and reduce mortality.

Another longitudinal study (Idler and Kasi, 1991) over a 4 year period reported mortality data for 2,812 participants in the Yale Health and Aging Project. For men reporting their health status as "poor" or "bad", the mortality rate was 6.75 times that of those reporting their health status as "excellent". The corresponding risk ratio for females was 3.12 and age was the best predictor of mortality in both men and women. For men, predictors of mortality were, in order, age, self assess health status, Roscow score, smoking and diabetes while for women, the predictors were age, diabetes, self assessed health status, Roscow score, body mass index and smoking. For women the association between mortality and self assessed health status existed only for those living in the community and not for those in public or private housing for the elderly. Medical care utilization, presence of social resources (close friends) and emotional resources (religiosity) had little or no direct effect on mortality in their study.

A Japanese study (Tsuji et al., 1994) noted self rated health status to correlate well with cancer mortality, functional (ADL) disability and stroke mortality while limitation in ambulatory activity significantly increased the risk of heart disease mortality.

Data collected by (Lyyra et al., 2006) in a 10 year follow-up Swedish study were similar to those collected in the Active Ageing dataset for an Australian population, however, they had access to all cause mortality data. Mortality for those with limited life satisfaction, with limited pleasure in everyday activities and without "meaning in life" was twice those in the highest quartile with respect to the Mood factor based on these and other depression related factors. A similar influence of mood was noted in a US study (Ostbye et al., 2006) which also noted vision and hearing to be significant predictors of overall self reported health while age, gender and cognition predicted survival. As had been shown by Svedberg et al. (2005), they considered the predictors of self reported health to be "potentially modifiable and amenable to clinical and public health efforts".

Bath (2003) also studied self reported health status as well as change in health status but was unable to show that these predicted mortality but his study showed social engagement to be an independent predictor of short and long term mortality. He suggested, therefore, that the impact of both physical and social activities and their impact on mortality should be studied in more detail.

The use of factor analysis in studies of ageing and mental health have been reported by Hsieh (2005) in his assessment of successful ageing in Taiwan and by Marquez et al (2006) in their reduction of a 30 item assessment for depression to a 5 item tool. Nybo et al (2001) were able to use factor analysis to identify 5 items of a 26 item activity of daily living (ADL) scale of functional assessment and to show a high correlation between the two.

The aim of this study was to assess the efficiency of various filter (subset and attribute), wrapper and clustering (SOM/Kohonen and K-Means) methods as means of feature reduction using a "real-world" dataset of a similar ageing population. Since there were no specific target attributes in this dataset, models using all attributes as targets were created and both supervised (wrapper and filter) and unsupervised (clustering) selection methods were used to define subsets upon which the models are based. Misclassification rates and measures of

complexity for these models were compared with those using the full dataset.

From a domain perspective the aim of the study was to assess the relevance of the various questionnaire segments in classifying the population studied. Subsequent studies of the Australian population could be more effective with a higher response rate if the questionnaire could be reduced from the 165 variables to one of more limited scope providing the reduction in variable count would not adversely affect the predictive value of the data. Thus, this attempt at assessing the effect of feature selection should be of benefit in designing subsequent studies of the needs of the population and may be of benefit in more efficiently determining factors predicting self reported health status in the Australian population. There was no intent to compare the validity of this questionnaire with that of the established questionnaires on which the Active Ageing questionnaire was based.

2 Method

The Active Ageing dataset is based on the 2,627 responses (46% response rate) to a 2004 questionnaire mailed to 6,000 members of the Australian population aged 50 years and over (Australian Active Ageing (Triple A) Study at Queensland University of Technology). Information sought in this survey was grouped into eight categories (Groups A-H) with these representing attributes shown in Table 1 below. This also shows the distribution of attributes in each group.

Table 1: Questionnaire Attribute Distribution

Attribute Group	Group Description	Attribute Count	Percent
A	Work	14	8.5
B	Learning	33	20.1
C	Social	11	6.7
D	Spiritual	9	5.5
E	Emotional	24	14.5
F	Health Home	50	30.5
G	Life Events	9	5.5
H	Demographics	13	8.5

The information sought in each section was based on several validated instruments for the assessment health status (Short Form-36v2 (Hawthorne et al., 2007), (Ware et al., 1993), (Sanson-Fisher and Perkins, 1998) and Visual Function Questionnaire (VFQ-25) (Mangione et al., 1998)), social support (based on Duke's Social Support Questionnaire) (Goodger et al., 1999), (Koenig et al., 1993) pychological well-being including the dimensions: autonomy, environmental mastery, personal growth, positive relations with others, purpose in life and self-acceptance (Clarke et al., 2001), built environment, learning (Purdie and Boulton-Lewis, 2003) and social life events (Life Events Questionnaire (LEQ) (Sarason et al., 1978)). The questionnaire used portions of each of the validated questionnaires. As well, segments on work, and volunteer activities, spirituality and demographics were included.

Of the returned questionnaires several were not analysed based on their being incomplete or completed by persons under age 50. The questionnaire included 178 variables but following initial pre-processing, the attribute set comprised 165 variables and it is these that are being analysed. The dataset contained redundant variables**.**

After simple pre-processing of the dataset, several wrapper, subset and attribute feature selection (FS) techniques as well as two clustering methods were applied to the dataset. The wrapper method used was "Wrapper" and the filter methods used were the subset evaluation methods, Cfs and Consistency. The search method used for these was BestFirst. Values for the merit of the chosen subset were available for Wrapper and Cfs methods. The attribute evaluation filter methods used were GainRatio, InfoGain, SymmetricalUncertainty, ReliefF, OneR and Chi-Squared and for these the search method used was Ranker. The clustering methods used were Self-Organising Maps / Kohonen (SOM/Kohonen) and K-Means (Dunham, 2003a) and the factors determined by the algorithm to be of importance in determining the contents of the clusters were used and the results are the cluster derived reduced subsets. K-Means clusters were determined using a squared error algorithm with the desired number of clusters being an input parameter. For the SOM/Kohonen clusters, competitive unsupervised learning of neural networks was used with weights being adjusted according to hidden features or patterns uncovered.

The wrapper and attribute feature selection experiments were conducted using the WEKA software (Witten and Frank, 2005) and the clustering experiments were conducted using the SAS software. Distribution of attributes in the reduced datasets as well as the misclassification rates (MR) and model complexity (rule count) were compared among FS methods. These results are presented for the total dataset since there were no clearly defined target attributes. It was for this reason, as well, that clustering, an unsupervised method, was used for feature reduction.

Misclassification rates for each of the models was calculated using J48, WEKAs implementation of Quinlan's C4.5 method (Quinlan, 1993). Model complexity here was based on the number of rules used in the classification model. Data partition of 90/10 was used for most analyses but for Chi-squared analyses, cross-validation was used. For several analyses both methods were used and there was not a significant, consistent difference noted (see Figure 8).

3 Results

3.1 Attribute Distribution

Attribute distribution amongst the various groups for each of the feature selection methods and all target attributes was assessed. Figures 1 and 2 show this distribution as a percentage of attributes comprising the subset that are selected from each of the feature groups and this distribution is compared with the percentage of attributes in each group for the total dataset (Table 1). These data reveal the subset evaluation method (Cfs) to display only a low to moderate concordance of attributes chosen in relation to the group of the target attribute with this ranging from 10..9% for Group G (Life Events) targets and subsets to 73% for Group F (Health and Home) subsets and targets.. Data for the Consistency method, as well, showed the concordance values ranging from 4.71% for Group G to 40.33% for Group B (Learning) attributes and targets with Group F concordance being only 16.27%. The concordance values for Cfs were highest for Groups E and F and moderate for Groups A, B and C (41.8,53.92,50.85% respectively) whereas for Consistency method Groups D, F, G and H concordance was low, while groups A, B, C and E were moderate at 36.22, 40.33, 39.42 and 26.32% respectively). With the attribute evaluation methods (GainRatio, InfoGain, SymetricalUncertainty, OneR and Chi-Squared) with the nominated "subset" size being 5, the concordance was much higher for Group F (ranging from 67.2% to 98%) and Group G (4.55 to 33.67%). GainRatio (Figure 2) was consistently at the higher end of the range and OneR consistently at the lower end. For all attribute evaluation methods, Groups A, B, C, E and F had moderate to high concordance rates. Group D (Spiritual) rates were in the 50-60% range.

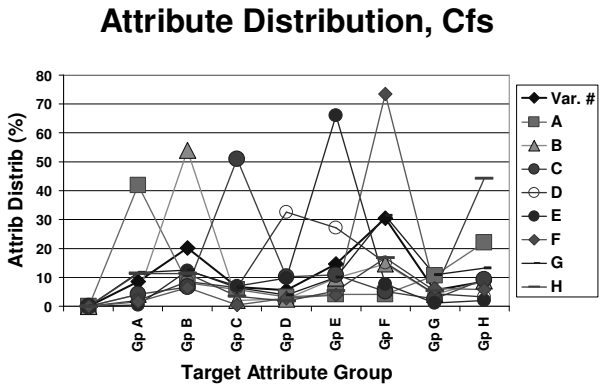

Figure 1: Attribute distribution in Cfs defined subsets (All targets, filter method, subset evaluation)

For all targets, each of the feature selection methods is better at choosing correlated features within the same group as the target attribute. Consideration of the content of subgroups for each of the target attributes reveals that in most cases there is domain significance associated with the attributes even when the selected attribute is in a different attribute group.

The subset evaluation feature selection methods result in subsets with attribute count ranging from 3 to 55 but most subsets contained between 5 and 30 attributes. There was less variation in the size of the Consistency derived subsets (counts mostly 7-11) whereas for the Cfs derived subsets the attribute count was usually higher and the range wider (3-55). For attribute evaluation methods in this study the count chosen for subset size was five.

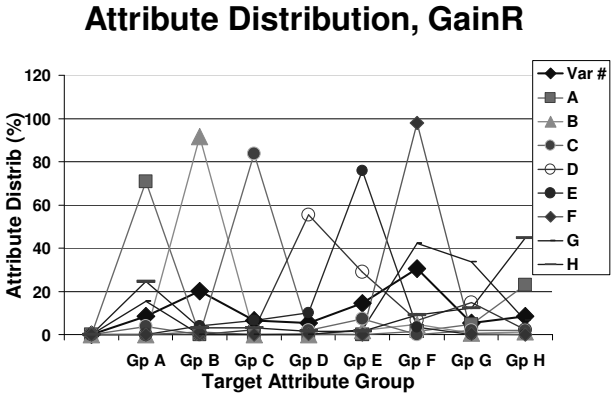

Figure 2: Attribute Distribution for GainRatio defined subsets (all targets, filter method, attribute evaluation)

The efficiency of each FS method in reducing the size of the dataset was assessed and subset methods, CFS and Consistency, were observed to be the least efficient while, for all other methods, 5 was the user-selected count. Compared to the full dataset with 165 attributes, this represents a reduction in attribute count of 63-97% with all attribute evaluation methods as used here yielding a reduction of 97%.

.

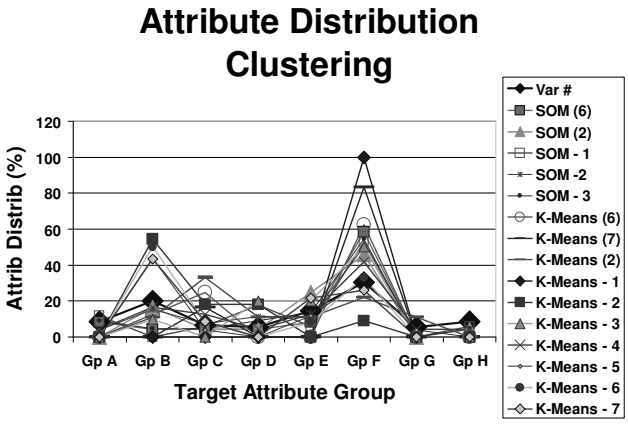

Figure 3: Attribute Distribution for clustering defined subsets

K-Means and SOM/Kohonen techniques were used to cluster the survey respondents and to assess the attributes with importance in determining cluster content. Figure 3 shows that the SOM/Kohonen technique consistently uses more attributes from Group F (Health and Home) and all have domain significance. On certain K-Means

experiments, the algorithm chose only a small number of attributes and compared to other K-Means analyses, these showed a bias to F group attributes while in the other analyses using K-means clustering, there is much greater use of Group B attributes. The Group B (Learning) attributes are much more likely to be from the "needs' and "wants" section of the learning attributes. The Group F attributes used in both clustering methods represent both physical and emotional characteristics of the population. Attributes from work, social, spiritual, emotional and demographic segments (Groups A, C, D, E and H) were used much less frequently.

For several runs of each clustering method, the average attribute count for SOM/Kohonen derived subsets was 22 and for K-Means 11.9 with the range being 4 – 25 represneting a reduction in attributes count in the order of 85-98%.

3.2 Misclassification Rate

Models using each attribute as a target attribute were assessed in hopes of more clearly defining what subset of features of this population are associated with valid predictive models. Misclassification rates for all target attributes using the reduced and full datasets were assessed. The results are shown in Figures 4 and 5. In general, there are limited differences among the various feature selection methods but for most target attributes, it is clear that the misclassification rate for the full dataset is higher than that for the models derived from the reduced datasets. Models based on work, health and home, life events and demographic targets are better predictors than those based on learning, social, spiritual and emotional characteristics of the population.

For Group A (paid and volunteer work) targets, the misclassification rate (MR) varies from 5% to over 60% (78.6% with MR less than 30%) while Group B (learning status, wants and needs) target models have MRs between 20 and 60% with only 27.3% models having MR of less than 30%. For Group C (social support) target models, the MR ranges from 25% to nearly 80% (36.4% with MR <30%) and for Groups D (spiritual) and E (emotional), there are no models for which the MR was below 30%. The MRs ranged from 32-66% for Group D and 31-64% for Group E. A large majority of Group F (health and home) and G (life events) target models have a MR of less than 30% (66% and 89% respectively) with Group F MR ranging 0.38 to 58.24% and Group G MR ranging 4.21-30.12%. For Group H (demographics), there are 76.9% of targets with model MR of less than 30% (range 3-69%).

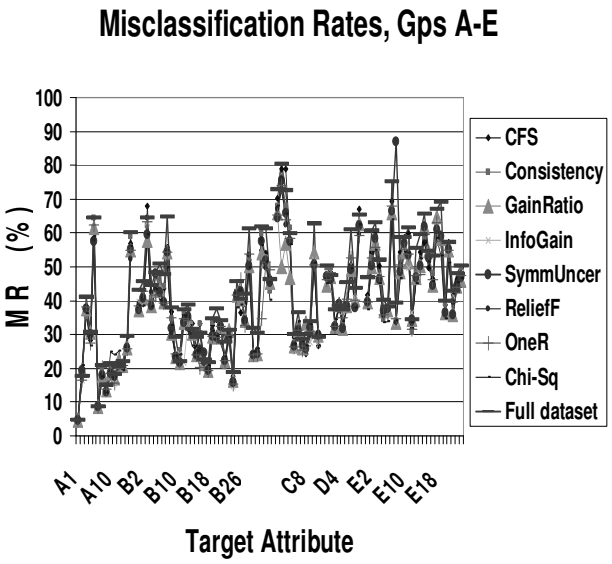

Figure 4: Misclassification rates for models associated with all target attributes (Groups A-E) using filter (subset and attribute) evaluation methods.

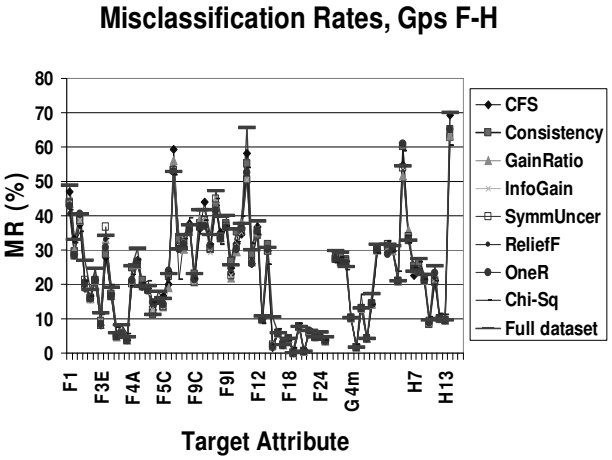

Figure 5: Misclassification rates for models associated with all target attributes (Groups F-H) using filter (subset and attribute) evaluation methods.

Figure 6 below shows the *percent* change in MR for models associated with all targets when compared with models derived using the full dataset. For most models built with reduced data set there is a 5-20% decrease in MR but for a relatively small number of models there is a decrement in accuracy with this being most marked for models derived from chi-square subsets.

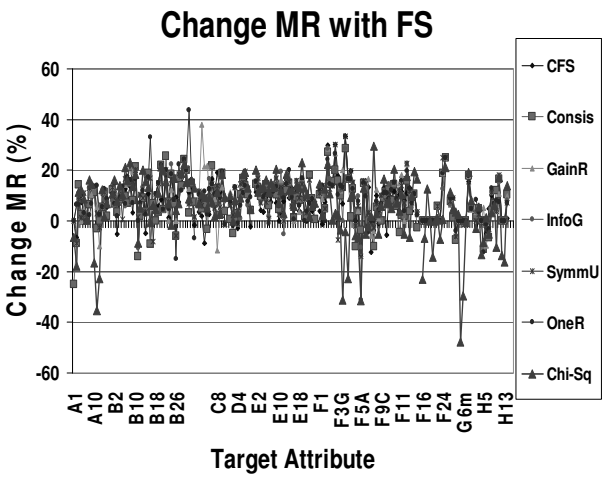

Figure 6: Change in Misclassification Rate with Feature Selection using all target attributes

Although not clearly shown in this composite figure, models with a misclassification rate of less than 30% are based on variables associated with paid work status and hours worked as well as volunteer work status and involvement in some course of study and on variables associated with learning limitation related to transport issues and lack of support, the social group variable (satisfaction with relationships) and several health and home variables (ability to engage in moderate activity, lift and carry groceries, climb several flights of stairs, walk more than a kilometre, limitation in type of work or activity, inability to accomplish planned tasks, activity limitation related to physical or emotional health, a feeling of being "down in the dumps" and ability to claim "my health is excellent"). Of the life events group of variables, those associated with highly predictive models include having suffered a personal illness or injury, undertaken new work or course of study, changed residence or suffered bereavement with the death of a spouse or partner. Demographic variables associated with strongly predictive models included being active in an organisation, having private health insurance and one's place of residence. For models associated with spiritual and emotional variables, none had a MR of less than 30% but those associated with an MR of less than 40% included a sense of being in control of one's life and contented as well as having a sense of direction and purpose and the ability to manage responsibilities of daily life.

Figure 7 shows the same data for selected models. Again, this shows that for almost all models associated with targets of domain importance there is an increase in model accuracy and models with reduced accuracy are mostly based on chi-square subsets.

Several targets shown to be associated with less predictive models in the above analyses were considered important in the determination of cluster membership using both SOM/Kohonen and K-Means.

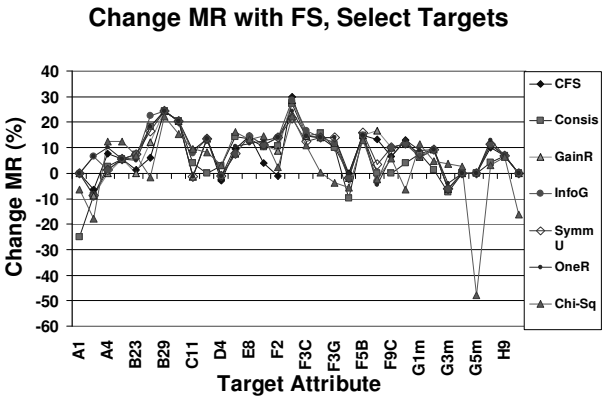

Figure 7: Change in Misclassification Rate with Feature Selection for selected target attributes (supervised methods)

K-Means method used more attributes from the learning "wants" and "needs" categories, such as needing to learn to manage own transport, to find people to manage one's finances and to trust with finances and to discourage violence as well as limitation to learning by lack of self confidence and the attitude of others. Personal contacts within one hours drive, frequent telephone contacts as well as a sense of being in control, contented, and happy were also used to determine cluster membership by the K-means method. Other K-means clustering attempts revealed a higher preponderance of health and home features to be of importance and these included ability to climb one flight of stairs, activity limitation related to physical health, feeling full of life. In one K-Means clustering attempt, F1 (self-reported health status) had an importance value of 0.28 compared to several learning characteristics with importance values of 0.56 to 1.0. Several factors related to mood and social activity scored higher than self reported health status. K-Means clustering, in general, found a broader range of features important in determining cluster membership but variables related to number of social contacts, various markers of mood and physical activity and an interest in learning were more consistently used.

SOM/Kohonen clustering consistently used more Health and home related variables and a very limited number of learning variables with these mostly being those relating limitation of learning by general health (B23). Emotional and mood related variables were frequently of high importance and markers of self assessed health ("health is excellent" (F11D), extent of activity limitation related to physical or emotional health (F6)) as well as other indicators of physical abilities, F3, F4, F5 components and mood related indicators from F9 components and variables from the social, spiritual and emotional groups.

Figure 8 below shows the MR for models associated with clustering derived subset target attributes (for both 90/10 data partition and cross validation using J48 as the classifier). These are plotted with the MR rates for the complete dataset using 90/10 data partition and J48 classifier.

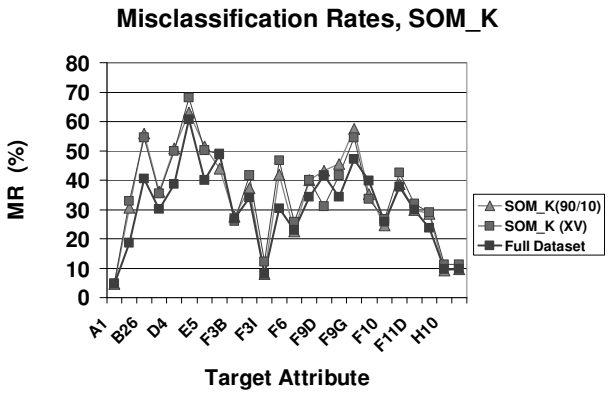

Figure 8: Misclassification rate for models associated with target attributes of SOM/Kohonen subsets (unsupervised methods). Comparison of MR as analysed using 90/10 data partition and cross-validation (XV).

Figure 8 shows there to be limited differences in the accuracy associated with use of the two validation methods and note is made of the use in SOM/Kohonen clustering of mostly Group F (Health and Home) attributes. Figure 9 below documents the change in MR when the SOM/Kohonen derived subset is compared to the MR for the full dataset and the same target attributes. It shows 15-65% increase in MR for many of the attributes. Both figures show clearly the reduction in accuracy associated with the use of SOM/Kohonen clustering as a means of feature reduction.

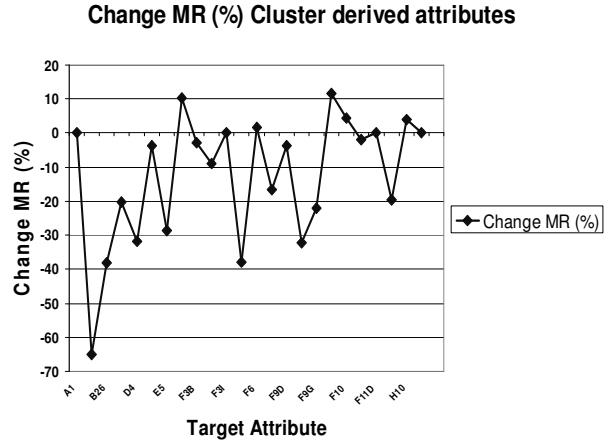

Figure 9: Change in Misclassification Rate for target attributes within a subset derived from SOM/Kohonen clustering

Evaluation of subsets as chosen by K-Means clustering is shown in Figure 10. The set of attributes for clustering with count of 7 and count of 6 overlapped and both are shown together with the MR for the full subset. As with subsets derived using SOM/Kohonen clustering, this method also identifies attributes, models for many of which result in MR that are higher than those using the full dataset. For only 1 of the 10 subsets was the MR for the reduced set improved on that of the full dataset and this was marginal.

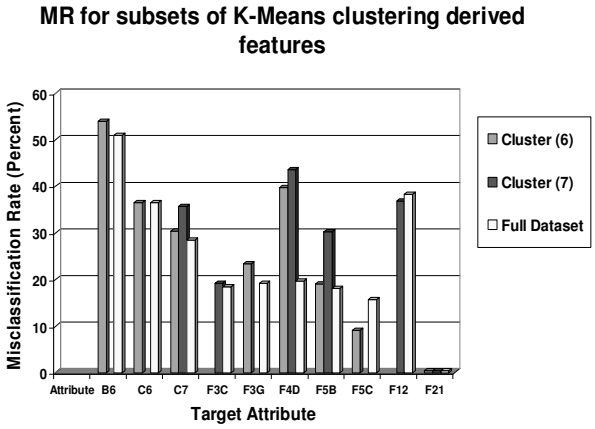

Figure 10: MR for K-Means derived subsets

Figure 11 shows the merit values for subsets derived by the "wrapper" methods and the subset evaluation method, Cfs. The "Wrapper" method assesses the predictive ability of each attribute individually as well as the degree of redundancy among them (Witten and Frank, 2005) and uses cross validation for accuracy assessment. The number of subsets evaluated ranged from approximately 800 to over 8,000 in the "Wrapper" and Cfs methods. The merit for subsets selected by "Wrapper" method are mostly higher than for those selected by Cfs but for target attributes F3 – F5 and F1-15, the Cfs merit values are higher. In the attributes relating to instrumental activities of daily living (F15-25), the merit values are low for both methods. As well, for attributes H10-12, the Cfs derived merit is higher than the "Wrapper" merit. This discordance may be related to the presence of "scarce attributes".

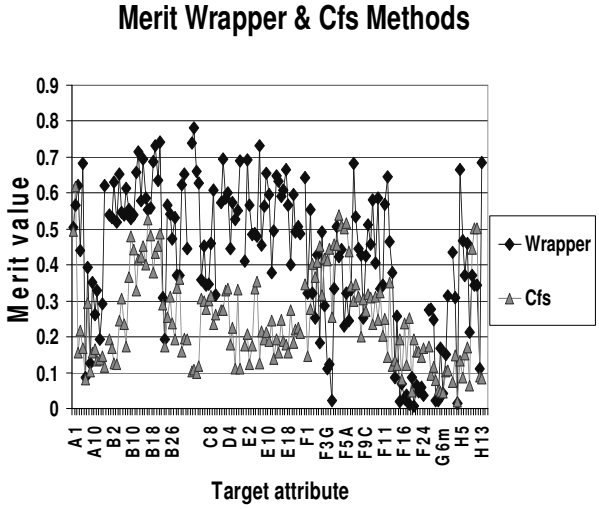

Figure11: Wrapper and Cfs merit values

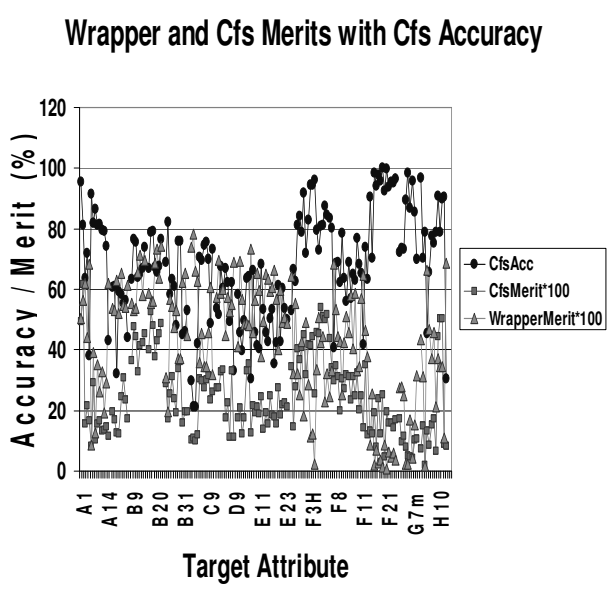

Figure 12: Comparison of Cfs Merit (as *percent)* and Accuracy of models derived from Cfs subsets

Figure 12 compares Cfs method and J48 classification to assess the accuracy of subset based models for all target attributes. This figure suggests that the filter method of subset evaluation when compared to the wrapper method overestimates the accuracy of models and that the wrapper method seems better able to address the issue of "scarce attributes", assigning a much lower merit to models associated with such scarce attribute targets..

3.3 Model Complexity

For all models derived in this study, the number of rules involved was recorded and figure 14 shows the rule counts while figure 15 shows the *percent* change in the number of rules in the model compared to that of the corresponding full dataset model. All of the attribute evaluation methods are able to reduce model complexity by 90-99% whereas the subset evaluation and clustering methods achieve a lower reduction in model complexity. Groups A, F, G and H have models most likely to contain no more than 20 rules and, for all groups the predictive models based on the reduced subsets, have significantly less rules than models based on the full dataset. The data points with zero change are mostly associated with models for which the feature selection methods could identify no correlated attributes.

Several attributes in the Health and Home, Life Events and Demographics groups were associated with models of one leaf only consistent with these attributes' being "scarce targets". It was for these models that the change in rule count, as shown in Figure 14, was zero.

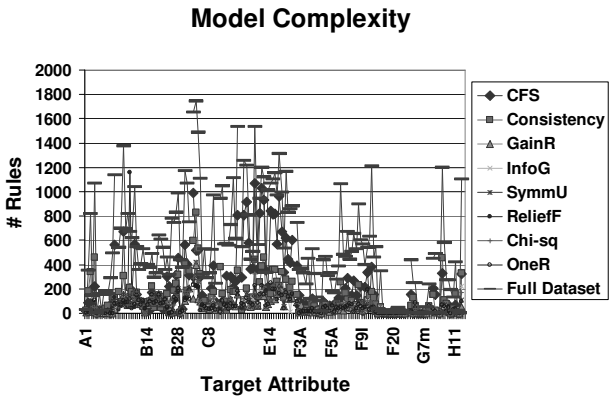

Figure 13. Model complexity for all targets

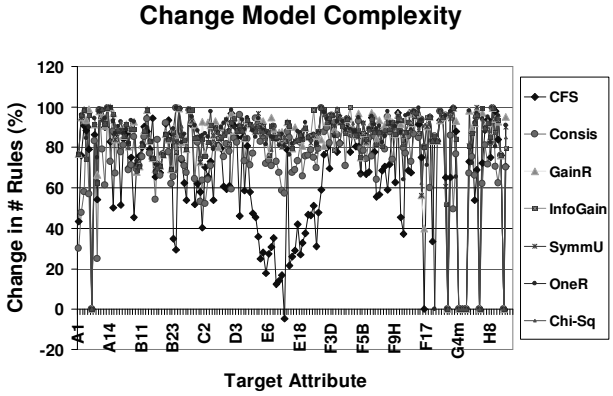

Figure 14. Change in model complexity with feature selection methods using all target models

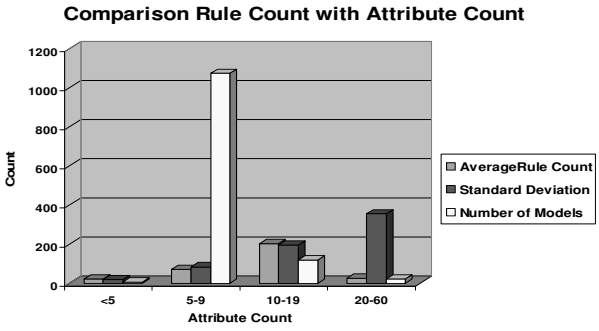

Figure 15: Comparison of Attribute count with Rule Count for all attribute and subset FS methods.

Figure 15 compares attribute count with rule count for models of all targets for all attribute and subset evaluation FS methods. This shows that for most models developed in this study, the attribute count is between 5 and 9 and for a relatively small increase in attribute count there is a significant increase in the rule count with marked increase in standard deviation.

Another marker of model complexity is the time for model development and validation. For subset and attribute evaluation methods, this was seconds and

frequently fractions of a second. The least time efficient model was the ReliefF method for which model construction took mostly 5-6.5 minutes and validation as long as 40 minutes for cross validation and 11 minutes for 90/10 data partition validation. ReliefF data for the full dataset is incomplete for this reason.

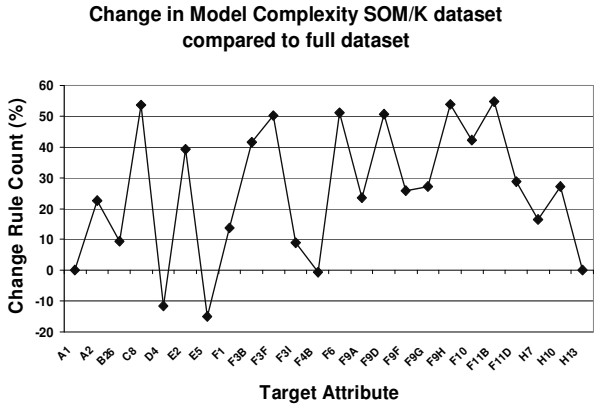

Figure 16: Change in model complexity in association with models for SOM/K derived subset

Figure 16 shows only a 10-55% reduction in model complexity for attributes of importance in defining SOM/K clusters compared to a mostly 60-97% reduction for the subset and attribute evaluation methods.

4 Discussion

The focus of this analysis of the Active Ageing Dataset was determination of the effect of several feature selection methods on the accuracy and the complexity of the models derived using the reduced datasets.

The results of the attribute distribution for each of the reduced datasets shows how the reduction in number of attributes in most cases increases the ability of the attributes in each subset to predict targets in the same attribute group. FS methods that are most effective in this goal are the attribute evaluation methods.

When considering the filter selection methods, the misclassification rate data show that there is no consistent loss of accuracy despite the reduction of the number of attributes in dataset to as little as 3% and mostly less than 15% of the attributes in the full dataset. The FS method resulting in significant worsening of model accuracy was Chi-Squared algorithm. The significance of the use of the cross validation as compared to data partition at 90/10 as was used for all other analyses is unclear but other analyses of many of the attributes using both methods did not show differences of the same magnitude. Using clustering techniques as a means of deriving a reduced dataset achieves a reduction to approximately 20% or less of the total number of attributes but there is a reduction in accuracy of many of the models derived from such datasets. The reason for this reduction in accuracy is unclear. The inclusion of a larger number of attributes results in larger models and this may contribute to the accuracy reduction. As well, the attributes of importance in clustering the population studied may be influenced by the somewhat homogenous nature of the population responding to the survey. This population has been shown to be still working or volunteering and mostly healthy (both physically and emotionally) as well as being of higher educational background and higher socio-economic class (Nayak and Buys, 2006), (Nayak et al., 2006). Under such circumstances, the importance of quality of life attributes in the social, emotional and spiritual groups may be somewhat overestimated in clustering of the full dataset.

The use of clustering algorithms for feature selection eliminates the need to nominate a target attribute. The variables important in clustering this population are consistently biased toward the Health and Home group. Of the other variables chosen in the SOM/Kohonen method, most have some domain significance and are consistent with the mood and instrumental activities of daily living features documented by others (Ostbye et al., 2006), (Lyyra et al., 2006). K-Means clustering algorithm is most useful when the number of clusters is known in advance (Witten and Frank, 2005) Since in this dataset the number of clusters is not known in advance, K-Means clustering may be less effective. The results obtained here show its strong bias to learning related attributes which have not been shown to be strong predictors of self reported health status or mortality. Clustering methods are however able to select attributes with obvious domain significance in an unsupervised manner and these attributes in many cases correspond with targets whose models in supervised learning methods have a high predictive accuracy. As well the attributes selected are frequently similar to those reported by other authors.

Models based on F1 (self assessed health status) as a surrogate for mortality (Mackenbach et al., 2002), (Nishisaki et al., 1996) have a lesser predictive value in this study than other related variables. The MR for the model associated with this attribute (44.6%) was higher than for many other variables in this dataset. Other markers of self assessed health such as F11D (“my health is excellent”) and F6 (extent of limitation of social activities by physical or emotional health) were associated with models of higher predictive accuracy (MR of 26.2% and 22.61% respectively). The failure to show a strong ability of F1 (self-assessed health status) in characterizing this population may be related to the fact that the respondents to this survey were mostly healthy and, in general, quite active physically and mentally. As well, most were living independently in their own home and were above the mean in background educational attainments and socio-economic status.

The subset and attribute evaluation feature selection methods do not differ significantly or consistently in their ability to select reduced datasets yielding models with at least no reduction in misclassification rate. In most cases the accuracy of the model prediction is increased. Complexity of the models is reduced more significantly in the attribute evaluation derived subset models than for the subset evaluation derived models. Clustering

methods studied here tend to provide datasets with a wider range of attributes and they result in models with reduced predictive accuracy for the component target attributes and model complexity greater than that for attribute evaluation derived subsets. The ability of the "wrapper" method to make allowance for scarce attributes seems to be a positive feature of this method as used for this dataset. Advantages of the wrapper methods have been noted as well by Kohavi and John (1997). This data did not show a significant difference in accuracy between models associated with attribute and subset evaluation methods as suggested by Yu and Liu (2003), however, the attribute methods did result in models with reduced complexity. The use of clustering methods to define subsets for analysis was able to reduce the complexity of models although not as well as was achieved by attribute evaluation methods and their use was associated with a reduction in predictive accuracy.

5 Conclusion

This study of the use of Feature Reduction algorithms and clustering in the assessment of a large population survey database has shown that the use of the subset and attribute evaluation methods mostly results in an improvement in accuracy of between 5 and 15% despite a reduction in the number of attributes assessed by 85-97%. Attribute evaluation FS methods consistently increase the accuracy of models despite a user chosen and maximal reduction in feature count while subset or wrapper evaluation methods use a larger number of attributes to achieve in many cases a lesser improvement in accuracy. Wrapper methods seem to allow for the presence of sparse targets in their analysis.

Choice of subsets using clustering methods allows assessment of the full dataset for which no definitive target attribute has been defined. It results, however, in only a moderate reduction in predictive model complexity (with a consequent broader range of attributes) and is associated with an increase in misclassification rate for many of the target attributes.. Cluster defined subsets result in a reduction in accuracy of the predictive model, however, clustering does allow selection of a range of attributes with domain significance and these attributes are frequently those with higher predictive accuracy when used as targets in supervised learning methods.

6 References

AHA, D. W. & BANKERT, R. L. (1995) A comparative evaluation of sequential feature selection algorithms *Fifth International Workshop on artificial intelligence and statistics.* Fort Lauderdale, FL.

BATH, P. (2003) Differences between older men and women in the self-rated health-mortality relationship. *Gerontologist,* 43**,** 387-95.

BENGIO, S. (2006) Statistical Machine Learning from Data Feature Selection. Matigny, Switzerland.

BHARGAVA, H. K. (1999) Data Mining by Decomposition: Adaptive Search for Hypothesis Generation. *INFORMS Journal on Computing,* 11**,** 239.

CLARKE, P. J., MARSHALL, V. M. & RYFF, C. D. (2001) Measuring psychological well-being in the Canadian Study of Health and Aging. *International Psychogenetics,* 13**,** 79-90.

DASH, M., & LIU, H (1997) Feature selection for classification. *Intelligent Data Analysis***,** 131-156.

DUNHAM, M. H. (2003a) Clustering. *Data Mining: Introductory and Advanced Topics.* Upper Saddle River, New Jersey, Prentice Hall, Pearson Education Inc.

DUNHAM, M. H. (2003b) *Data mining introductory and advanced topics*, Upper Saddle River, NJ : Prentice Hall/Pearson Education.

GOODGER, B., BYLES, J., HIGGANBOTHAM, N. & MISHRA, G. (1999) Assessment of a short scale to measure social support among older people. *Australian and New Zealand Journal of Public Health,* 23**,** 260-265.

HAWTHORNE, G., OSBORNE, R., TAYLOR, A. & SANSONI, J. (2007) The SF36 Version 2: critical analyses of population weights, scoring algorithms and population norms. *Quality of Life Research,* 16**,** 661-673.

HSIEH, C.-M. (2005) Age and relative importance of major life domains. *Journal of Aging Studies,* 19**,** 503.

IDLER, E. L. & KASI, S. (1991) Health perceptions and survival: do global evaluations of health status really predict mortality. *J Gerontol.,* 46**,** 555-65.

IDLER, E. L., KASI, S. V. & LEMKE, J. H. (1990) Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982-1986. *American Journal of Epidemiology,* 131**,** 91-103.

KAWADA, T. (2003) Self-rated health and life prognosis. *Arch Med Res,* 34**,** 343-7.

KOENIG, H. G., WESTLAND, R. E., GEORGE, L. K., HUGHES, D. C., BLAZER, D. G. & HYBELS, C. (1993) Abbreviating the Duke Social Support Index for Use in Chronically Ill Elderly Individuals. *Psychosomatics,* 34**,** 61-69.

KOHAVI, R. & JOHN, G. H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence,* 97**,** 273-324.

LANGLEY, P. (1994) Selection of relevant features in machine learning. *AAAI Fall Symposium on Relevance.*

LEE, Y. (2000) The predictive value of self assessed general, physical, and mental health on functional decline and mortality in older adults. *J Epidemiol. Community Health,* 54**,** 123-9.

LYYRA, T.-M., TÖRMÄKANGAS, T. M., READ, S., RANTANEN, T. & BERG, S. (2006) Satisfaction With Present Life Predicts Survival

in Octogenarians. *The Journals of Gerontology,* 61B**,** P319.

MACKENBACH, J., SIMON, J., LOOMAN, C. & JOUNG, I. (2002) Self-assessed health and mortality: could psychosocial factors explain the association? *Int J Epidemiol,* 31**,** 1162-8.

MANGIONE, C. M., LEE, P. P., PITTS, J. & AL, E. (1998) Psychometric properties of the National Eye Institute Visual Function Questionnaire, the NEI-VFQ. *Arch Ophthalmol.,* 116**,** 1496-1504.

MARQUEZ, D. X., MCAULEY, E., MOTL, R. W., ELAVSKY, S. & AL, E. (2006) Validation of Geriatric Depression Scale-5 Scores Among Sedentary Older Adults. *Educational and Psychological Measurement,* 66**,** 667.

MAXSON, P. J., BERG, S. & MCCLEARN, G. (1996) Multidimensional patterns of aging in 70-year-olds: Survival differences. *Journal of Aging and Health,* 8**,** 320.

NAYAK, R. & BUYS, L. (2006) Data Mining in Conceptualising Active Ageing. *Australian Data Mining Conference.* Sydney.

NAYAK, R., BUYS, L. & LOVIE-KITCHIN, J. (2006) Influencing Factors in Achieving Active Ageing. *Workshop on Optimisation based Data Mining Techniques with applications, ICDM.* Hong Kong.

NISHISAKI, S., UTOGUCHI, K., MIZOUE, T., TOKUI, N., OGIMOTO, I., IKEDA, M. & YOSHIMURA, T. (1996) The association of self-rated health and mortality -- a 7 year follow-up study of a Japanese community (English abstract only).

NYBO, H., GAIST, D., JEUNE, B., MCGUE, M., VAUPEL, J. & CHRISTENSEN, K. (2001) Functional status and self-rated health in 2,262 nonagenarians: the Danish 1905 Cohort Survey. *J Am Geriatric Soc,* 49**,** 601-9.

OSTBYE, T., KRAUSE, K., NORTON, M. C. & TSCHANZ, J. (2006) Ten Dimensions of Health and Their Relationships with Overall Self-Reported Health and Survival in a Predominately Religiously Active Elderly Population: The Cache County Memory Study. *Journal of the American Geriatrics Society,* 54**,** 199.

PURDIE, N. & BOULTON-LEWIS, G. (2003) THE LEARNING NEEDS OF OLDER ADULTS

THE LEARNING NEEDS OF OLDER ADULTS. *Educational Gerontology,* 29**,** 129.

QUINLAN, R. (1993) *C4.5: Programs for Machine Learning,* San Mateo, CA, Morgan Kaufman.

SANSON-FISHER, R. W. & PERKINS, J. J. (1998) Adaptation and validation of the SF-36 health survey for use in Australia. *Journal of Clinical Epidemiology,* 51**,** 961-967.

SARASON, I. G., JOHNSON, J. H. & SIEGEL, J. M. (1978) Assessing the impact of life changes: Development of the life experience survey. *Journal of Consulting and Clinical Psychology,* 46**,** 932-946.

SVEDBERG, P., GATZ, M., LICHTENSTEIN, P., SANDIN, S. & PEDERSEN, N. L. (2005) Self-Rated Health in a Longitudinal Perspective: A 9-Year Follow-Up Twin Study. *The Journals of Gerontology,* 60B**,** S331.

TANG, W. & MAO, K. (2005) Feature Selection Algorithm for Data with Both Nominal and Continuous Features. *Advances in Knowledge Discovery and Data Mining.* Springer Berlin / Heidelberg.

TSUJI, Y, M., PM, K., S, H., H, A., M, S. & K, S. (1994) The predictive power of self-rated health, activities of daily living and ambulatory activity for cause-specific mortality among the elderly: a three-year follow-up in urban Japan. *J Am Geriatr Soc,* 42**,** 153-6.

WARE, J. E., SNOW, K. K., KOSINSKI, M. & GANDEK, B. (1993) *SF-36 health survey: Manual and interpretation guide,* Boston, The Health Institute, New England Medical Centre.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.

YU, L. & LIU, H. (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03).* Washington, D.C.