Constructing a Synthetic Longitudinal Health Dataset for Data Mining

Shima Ghassem Pour*, Anthony Maeder* and Louisa Jorm†
*School of Computing, Engineering and Mathematics
University of Western Sydney, Campbelltown, Australia
Email: A.maeder@uws.edu.au
†School of Medicine, University of Western Sydney, Campbelltown, Australia
Email: L.jorm@uws.edu.au

***Abstract*—The traditional approach to epidemiological research is to analyse data in an explicit statistical fashion, attempting to answer a question or test a hypothesis. However, increasing experience in the application of data mining and exploratory data analysis methods suggests that valuable information can be obtained from large datasets using these less constrained approaches. Available data mining techniques, such as clustering, have mainly been applied to cross-sectional point-in-time data. However, health datasets often include repeated observations for individuals and so researchers are interested in following their health trajectories. This requires methods for analysis of multiple-points-over-time or longitudinal data. Here, we describe an approach to construct a synthetic longitudinal version of a major population health dataset in which clusters merge and split over time, to investigate the utility of clustering for discovering time sequence based patterns.**

***Index Terms*—cluster analysis; synthetic data**

I. Introduction

The rapid growth of digital information and the related capability of collecting and storing huge amounts of data offer two new opportunities. On one hand, these data represent the potential for discovering useful information and knowledge which has not been uncovered before because of the sheer volume of data which is now available. On the other hand, the limited ability of conventional processing of large amounts of data to discover useful information and new knowledge can be overcome by applying a new generation of mathematical techniques to extract patterns from the data [1]. Data availability is increasing exponentially, while the human level of processing ability is almost constant. As this gap increases, there is a growing necessity for knowledge discovery in databases and data mining [1], [2].

Data mining refers to discovering insight from data, which is reliable statistically and not known as priori [3]. This data must be available, relevant, adequate, and clean. Also, the data mining problem must be well-defined, yet cannot be solved simply by query and reporting tools, and should be guided by a data mining process model [4]. Overall, data mining is the process of discovering and interpreting previously unknown patterns in databases [5]. There are many methods of data mining used for different purposes and goals.

We can use data mining for discovering structure in large volumes of data. Most data mining techniques aim to analyse only one single set of data elements which are static in time. But, often in health data, we have many sets of observations at different times and we want to follow their trajectories over time. Longitudinal data is essentiality data observed sequentially over time [6]. It may be collected from a designed experiment or an observed study, where the outcome variables are related to a sequence of event or responses recorded at certain time point during study period. In particular, large scale longitudinal data is frequently encountered in research in biology, medicine, and public health.

The organization of the paper is as follows: in Section II we describe the problem of analysing a longitudinal dataset, in Section III we present the proposed method for creating a synthetic dataset, and in Section IV we discuss experimental result using the synthetic dataset.

II. Problem

Our overall research addresses the area of cluster analysis and attempts to reformulate cluster analysis for application to longitudinal health data problems, with a resulting improvement in conceptual simplicity and statistical capability. In clustering methodology, one is generally given a sample of N objects, each of which is measured on M variables. From this information alone, one must devise a classification scheme for grouping the objects into K classes. The number of classes and the characteristics of those classes are to be determined to solve some underlying problem.

If all the objects in a given class were identical to one another, the problem would be simple. However, in the usual situation the objects in each class differ on most or all of the measures [2]. Most cluster analysis procedures try to measure the similarity between any two objects, and then try to group the objects so as to maximize within-class similarity. Unfortunately, the appropriate measure of similarity is a subject of some controversy. It would be desirable to derive a cluster analysis system without arbitrary assumptions about similarity [7], [8].

Since the objects within a class differ from one another, it is reasonable to assume the existence of probability distributions of characteristics for a population belonging to this class. Lazarsfeld and Henry introduced Latent Structure Analysis

[9], which is closely related to the mixture analysis problem. In "Latent Class Analysis" [10] (which is a special case of Latent Structure Analysis) the associations among variables are explained by assuming the population is a mixture of potential or 'latent'classes, within each of which the variables are independently distributed.

Latent Class Analysis (LCA) methods classify subjects into one of K unobserved classes based on the observed data, where K is a constant and known parameter. These latent or potential classes are then refined based upon their statistical relationships with the observed variables. LCA is a probabilistic clustering approach: although each object is assumed to belong to one cluster, there is uncertainty about an object's membership of a cluster. This type of approach offers some advantages in dealing with noisy data or data with complex relationships between variables, although as an iterative method there is always some chance that it will fail to converge.

An important related issue in how to make use of clustering methods such as LCA is to decide on the ideal number of clusters. Researchers use a range of information criteria to determine the number of clusters which best fit the data, such as Akaike Information Criterion (AIC) [11] and Bayesian Information Criterion (BIC) [12]. BIC is known as a good indicator for making a decision about the number of clusters for large datasets [13]–[15]. If L is the maximized value of the likelihood function for the estimated model, n represents the sample size and P indicates the number of parameters for the estimated model, BIC is represented by [12], [16]:

$$BIC = -2logL + Plog(n) \tag{1}$$

Longitudinal datasets are created when the same characteristics are measured repeatedly over time and can take a long time to build up if the time steps are numerous and of long duration. Complex health datasets contain different types of variables such as demographics, lifestyle and health measures, which can be difficult to analyse without long time sequences.

To understand the workability and stability of data mining methods on such data, there is a need to have an alternative way to create a longitudinal dataset. In this paper, we describe an approach to construct a longitudinal dataset synthetically. Use of a synthetic dataset would not be as accurate as using existing datasets, however a synthetic dataset would share some of properties of interest that one would expect to see in actual longitudinal datasets.

III. Method

The 45 and Up study [17] is a population-based cohort study with participants aged over 45 resident in the Australian State of New South Wales, randomly selected from the Australian Medicare database [18]. The study is organized by the Sax Institute and State Government of New South Wales as core partners and also has some partners from public health and health service research centres and universities across New South Wales [19], [20]. Recruitment into the 45 and Up study began in February 2006: by July 2008 the first 103,042 participants had joined the study [19] and since then it has more than doubled. The 45 and Up study combines socio-economic and demographic factors, as well as health and lifestyle. It is especially useful for studying effects of slow and chaotic emergence of diseases such as cancers.

The 45 and Up database serves as a good foundation dataset for testing data mining methods because of the richness of these different kinds of variables contained in it. Also it will in the future serve as a benchmark time based dataset, recording trajectories of individual health histories over a long period of time (potentially from age 45 until death). At the present time it is entering the second cycle of data collection providing a time step of approximately 2 years for most participants who have remained part of the study. Consideration of how longitudinal data mining approaches could be developed for use on this database in the longer term would be a useful contribution to extracting further value from it. For example, we are investigating a subset of the data concerning prostate cancer prevalence and risk factors and requiring longitudinal analysis.

We will construct a synthetic time version of data to test the ability of LCA to discover time sequence based patterns in this data. The synthetic time data will be formed by identifying a group of variables which could reasonably lead to splitting and merging of clusters over successive time steps. A simple example in the area of lifestyle related data is changes in body mass index (BMI), and exercise or Physical Activity (PA). These two variables were chosen as they are deemed most influential of those available.

Data from the first (current) time step will be clustered according to these variables, and then a revised set of data for the next successive time steps will be calculated using the above identified variables to produce converging and diverging sets of values. Addition of some controlled noise or perturbation of these calculated values according to an anticipated variation pattern could also be incorporated. As the imposed structure will be known in order to carry out these operations, this synthetic data can serve as a kind of gold standard for validating the clustering methods we will be applying. In order to construct a synthetic time version of 45 and Up dataset, we are interested to group the data based on variations of BMI and PA over time.

To create a synthetic dataset we follow these 5 steps:

Step 1: The first step is to choose a sample dataset from the actual dataset (in this case the 45 and Up dataset). We randomly select 23,000 elements to create a reasonably large dataset. For this sample dataset we choose the two variables, computed body mass index (BMI) and how many times per week subjects do some physical activity (PA). We have a BMI range from 16-44 and PA range from 0-100 times per week. To simplify the situation we divide the BMI variable into 4 different categories: the first category is people with BMI<18.5 (underweight), the second category people with BMI between 18.5 to 25 (normal), the third group people with BMI between 25 to 30 (overweight) and finally the last category with BMI>30 (obese). We divide the PA variable into 10 categories, people who have PA<10 times a week belong to the first category, PA between 10-20 in the second category and so forth. As a conclusion at this step we have 23,000 cases with 2 different variables: BMI variable with four categories and PA variable with 10 categories.

Step 2: In the second step, we apply LCA to cluster our data based on BMI and PA, and the result shows that two clusters fit our baseline data best (under the BIC criterion). People who are underweight, normal weight and overweight belong to Cluster1_B (52.6%) while Cluster2_B consists mostly of obese people (47.4%).

Step 3: Next, we change some of the values for the above two variables to vary how elements from the two clusters that we have from the first time step will appear in the second time step. To achieve this, we decide to choose people who were in Cluster1_B and randomly choose about one third of them to remain as before (3,797 cases), and divide half of the rest of them in two groups. For the majority (2,249) of them we change the BMI to overweight and the rest of them (1,519) we move to the normal BMI group. The other half of the BMI range is unchanged, however the PA is changed for this group: we randomly choose about one third of them (1,265) and increase the amount PA and decrease PA for the rest of them (2,531).

Step 4: The next step is to change Cluster2_B of the first time step. We split Cluster2_B in two groups equally. The first group remains without any change while the other is divided in two new groups. The first new group includes 4,212 cases which are 1,054 cases of decreased PA and 3,160 increased PA. In the second new group which includes 2,106 cases we change the BMI in 1,264 cases to the underweight category and the rest to the obese category.

Step 5: The final step is to apply LCA to the second time step to see how two clusters in the first time step actually split under the same clustering method as was applied to the first time step data.

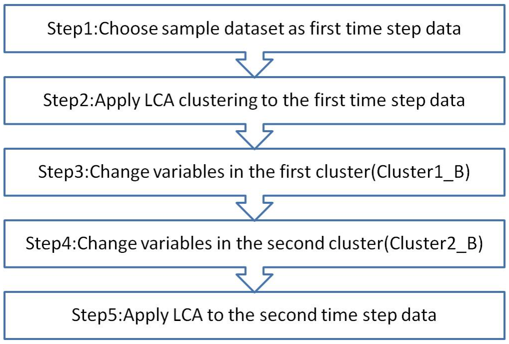
Step1:Choose sample dataset as first time step data
Step2:Apply LCA clustering to the first time step data
Step3:Change variables in the first cluster(Cluster1_B)
Step4:Change variables in the second cluster(Cluster2_B)
Step5:Apply LCA to the second time step data

Fig. 1. Five step creation of synthetic dataset

This sequence of steps is summarized in Fig. 1. At each step in the synthesizing we also have the opportunity to add some controlled noise to each group, but in the interests of simplicity we have not included that aspect in the work reported here. We can also repeat the above steps to construct more synthetic data for another time step. Our synthetic dataset will give us a good understanding of splitting and merging clusters over time as some of groups are designed differently but with some similarity to force the cluster to merge in the next time step. Additionally, when future time step data from 45 and Up becomes available, the methods we developed can be applied to that and a comparison can be made as to the similarity with our simulated results. For this aspect of the work, a minimum of 3 time steps will be needed in the dataset.

IV. Results

For this research project we used the LatentGOLD software package to perform LCA clustering due to the attractive range of additional features, and its widespread user base as noted from our readings in the literature. LatentGold is a commercial product [21] that uses SPSS to automatically provide a variety of output, graphics and diagnostics to help the user interpret the resulting clusters and to refine their analysis.

Applying Latent Class Analysis to the above synthetic dataset shows us that at the baseline our data fitted well into a two cluster model, with the majority of people who are obese but with different amounts of physical activity belonging to Cluster1_B and the rest of people in Cluster2_B. Table 1 shows how the clustering using LCA can be compared over four different models with respectively K = 1 to 4. The parameter BIC is an estimate of the overall tightness or stability of the clusters, while Npar is the number of free parameters to be estimated, LL is the log-likelihood and L2 is the square likelihood value for the estimated model. As the BIC values show, the model with 2 clusters is optimal for explaining the data. Fig. 2 shows the range of baseline data elements at the first time step assigned to Cluster1_B (circles) which includes mostly people who are underweight, normal and overweight, and Cluster2_B (crosses) which is exclusively people who are obese.

TABLE I
FOUR DIFFERENT MODELS FOR BASELINE DATA

K	LL	BIC(LL)	Npar	L2
1	-56492.24	113105.58	12	498.05
2	-56261.70	112775.70	25	36.97
3	-56250.99	112885.47	38	15.55
4	-56243.54	113001.78	51	0.66

Table 2 shows that for the synthetic data for the second time step, there are 4 clusters which best fit the data according to the BIC measure, Fig. 3 shows how these clusters cover the data space, with some elements of Cluster1_B from the first time step now moved in to the expanded Cluster2_B which becomes Cluster4_S in the second time step. The other elements from Cluster1_B in the first time step are split across 3 new clusters which roughly map to underweight (Cluster1_S), normal (Cluster2_S) and overweight (Cluster3_S) people. It is noted from Fig. 2 that the split of the first time step Cluster1_B is not entirely dependent on the BMI variable.

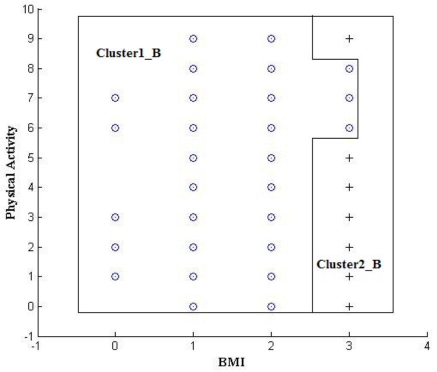

Fig. 2. Two cluster explanation of baseline data

The underweight cluster (Cluster1_S) extends to include some normal people with low exercise values, and the overweight cluster (Cluster3_S) extends to include some normal and underweight people with high exercise values (144 cases in all).

TABLE II
SIX DIFFERENT MODELS FOR SYNTHETIC DATA

K	LL	BIC(LL)	Npar	L2
1	-57679.5	115470.0	11	35529.41
2	-43022.1	86276.10	23	6214.50
3	-40551.0	81454.98	35	1272.36
4	-39916.0	80306.04	47	2.40
5	-39915.7	80426.31	59	1.66
6	-39915.7	80547.34	71	1.67

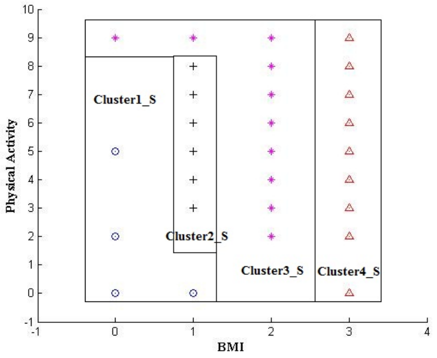

Fig. 3. Four cluster explanation of synthetic data

V. CONCLUSION

Analysis of longitudinal data is becoming popular as researchers are more interested to describe how people change during time. However, collecting an appropriate longitudinal dataset takes a long time to build up. On the other hand, to validate the workability and stability of data mining methods, there is need to have such a dataset available. The work here was done to explain how we can construct a longitudinal dataset synthetically. Such synthetic datasets will give us better understanding of methods to extract more useful and valuable information from our data, to explain how characteristics of people's health will change during time. In order to create a synthetic dataset, we followed a systematic approach which was based on the use of LCA to understand structural characteristics of the variables. Based on the initial clustering, we decided which variables should be changed and how this change should be applied, in order to simulate a desired change in the characteristics. This approach was tested using the 45 and Up dataset. These results shown here confirm that the desired effect was observed and the approach did not adversely affect clustering stability for the chosen example. We are now proceeding with creation of further large scale synthetic datasets for prostate cancer pattern analysis for multiple variables over multiple time steps.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan Kaufmann, 2011.

[2] H. Mannila, "Data mining: machine learning, statistics, and databases," in *Scientific and Statistical Database Systems, Proceedings*. IEEE, 1996, pp. 2–9.

[3] C. Elkan, "The foundations of cost-sensitive learning," in *International Joint Conference on Artificial Intelligence*, vol. 17, no. 1, 2001, pp. 973–978.

[4] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski, "Subgroup discovery with cn2-sd," *The Journal of Machine Learning Research*, vol. 5, pp. 153–188, 2004.

[5] D. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. The MIT Press, 2001.

[6] D. Hand, "Statistics and data mining: intersecting disciplines," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 1, pp. 16–19, 1999.

[7] C. Fraley and A. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

[8] ——, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[9] P. Lazarsfeld and N. Henry, *Latent structure analysis*. Houghton, Mifflin, 1968.

[10] L. Collins and S. Lanza, *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons Inc, 2009, vol. 718.

[11] H. Akaike, "Factor analysis and aic," *Psychometrika*, vol. 52, no. 3, pp. 317–332, 1987.

[12] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[13] L. Collins, P. Fidler, S. Wugalter, and J. Long, "Goodness-of-fit testing for latent class models," *Multivariate Behavioral Research*, vol. 28, no. 3, pp. 375–389, 1993.

[14] J. Hagenaars and A. McCutcheon, *Applied latent class analysis*. Cambridge Univ Pr, 2002.

[15] J. Magidson and J. Vermunt, "Latent class models," *The Sage handbook of quantitative methodology for the social sciences*, pp. 175–198, 2004.

[16] K. Nylund, T. Asparouhov, and B. Muthén, "Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study," *Structural Equation Modeling*, vol. 14, no. 4, pp. 535–569, 2007.

[17] "Study overview," May 2010. [Online]. Available: http://www.45andup.org.au/[accessedOct2011]

[18] N. Mealing, E. Banks, L. Jorm, D. Steel, M. Clements, and K. Rogers, "Investigation of relative risk estimates from studies of the same population with contrasting response rates and designs," *BMC Medical Research Methodology*, vol. 10, no. 1, pp. 10–26, 2010.

[19] E. Banks, L. Jorm, S. Lujic, and K. Rogers, “Health, ageing and private health insurance: baseline results from the 45 and up study cohort,” *Australia and New Zealand health policy*, vol. 6, no. 1, pp. 6–16, 2009.

[20] E. Banks, S. Redman, L. Jorm, B. Armstrong, A. Bauman, J. Beard, V. Beral, J. Byles, S. Corbett, R. Cumming *et al.*, “Cohort profile: the 45 and up study.” *International Journal of Epidemiology*, vol. 37, no. 5, pp. 941–947, 2008.

[21] “Latent gold,” May 2011. [Online]. Available: http://www.statisticalinnovations.com/products/latentgold.html[accessedMarch2011]