http://doi.org/10.4038/kjm.v6i1.7523

Analysing Sri Lankan Lifestyles with Data Mining: Two Case Studies of Education and Health

Mohotti W. A.[1], Premaratne S.C.[2]

[1]Faculty of Science, University of Ruhuna, Matara, Sri Lanka,

[2]Faculty of Information Technology, University of Moratuwa,

Moratuwa, Sri Lanka

[1]mohottiwathsala@yahoo.com,

[2]samindap@uom.lk

Abstract

There are no adequate researches in lifestyle data analysis of Sri Lanka. Existing works are not capable of handling big data systematically and, not efficient in disclosing the latent factors in lifestyles. This research has used predictive and descriptive mining techniques to analyse HIES dataset related to two cases in education and health using WEKA and SPSS. The design consists of a classification and a clustering task. Classification finds the factors and their relationships associated with no-schooling and dropouts. Clustering explores the relationship between chronic diseases and family income. Our analysis reveals the significance of child labor and religion among the influences such as age, district and parental education related to school dropout and no-schooling. Other case study discloses that citizens who have got low income comparatively suffer from many chronic diseases. The important patterns recognized through the research can be used by the government policy makers to setup policies.

Keywords: Citizens lifestyle, Classification, Clustering, Government Policies

Introduction

Lifestyle of a citizen is a composition of motivations, needs, and wants which is influenced by factors such as culture, family, reference groups, and social class. According to existing literature, well established European countries focused on factors such as income, health, education, labour market, household structure and living conditions to determine the living style of Europe. Sri Lanka also pays attention to same factors such as education, health, and

http://orcid.org/0000-0002-5720-7737

housing in determining development of citizen's life style (Smith, 1999). Socio-economic segmentation of their citizens based on these features is key concern area for a government to set appropriate future plans to uplift citizens' life quality. On the other hand, pay attention to real patterns and trends reveal by the citizens lifestyle can be used to find out misuse of government funds and irregularities.

But analysing useful factors for better decision making is a problem due to huge amount of data generated and associated with lifestyles. Also lack of initiatives and awareness in higher authorities for systematic in-depth analysis is another shortcoming in citizen lifestyle analysis. Though Central bank and Department of Census and Statistics collect household data related to lifestyles, those institutes are not paying attention to hidden pattern behind them. In most of the currently existing researches in the world, these factors have been examined carefully using principal component analysis and clustering techniques to identify their dominant combinations for activity patterns (Jiang et al., 2012). Then those resulted association rules are used to improve citizen satisfaction. But in Sri Lanka, those data are not effectively use to find out the socio-economic issues. If properly analysed and investigated those gigantic life style data which have been kept in government institutes, governance of the country could be improved and controlled. Those analyses can be used to track the areas which government should focus to uplift the life quality.

Therefore the research proposes a method for analysing lifestyle in Sri Lanka to investigate the hidden issues. This analysis is supported by information extracted from a Household Income and Expenditure survey (HIES) of Department of Census and Statistics (Department of Census and Statistics, 2013). Predictive and Descriptive data mining methods, i.e., classification and clustering used to discover linkages of household demographics and citizens' behaviour patterns.

Literature Review

Most of the international level household survey focuses on income and expenditures. The main aim of analysing those household survey data is to prepare estimation for fields such as education, health, transportation and communication. Furthermore, insight about the issues such as types of education perceived by household members or cost and benefits of health services used by family members, can be figured out through them.

In Europe, major areas that are focused for the socio-economic statistics include income, living conditions of Europe's households, health, labour market and education. By analysing those Europe household data, those countries try to find answers for critical questions such as how many households considered "at–risk-of poverty", presence of longstanding illness or disability, distribution of labour earnings and provision of public services. Furthermore, many Asian countries of the developing world highlight the need for deeper thoughtfulness of not simply the numbers of the poor but also the nature of poverty. Hence those countries depend on fields such as normal demographics data, income, education, occupation and location of residence. According to Asian bank report, most of the Asian countries are mainly focus on income, education, health, agriculture and other social protection schemes to improve Asians' life standards. As mentioned in central bank report main focus areas of Sri Lanka to determine the lifestyle are communication, power consumption, transportation, water supply and irrigation, education, health sector, housing, urban development and environment (Central Bank of Sri Lanka, 2013).

Department of census and statistics is a government authority which use to collect and analyse data to determine development of Sri Lanka. Household Income and Expenditure Survey (HIES) is one survey conducted by them to provide information on household income and expenditure to measure the levels and changes in living conditions of the people. Data collected from this survey is used to observe the consumption patterns to compute various other socio-economic indicators such as poverty-price indices. The HIES questionnaire consider nine priority areas to collect household information covering the demography, school education, health, food and non-food expenditure, income, inventory of durable goods, access to facilities in the area and debts of the households, housing information and agriculture holdings & Livestock. These considered factors are aligned with priority areas of central bank report (Department of Census and Statistics, 2013). Though department of census and statistics explores HIES data set to find out important economic and social indices, it does not pay attention to find out hidden pattern behind those data.

Lifestyle data of citizens in a country will be huge in size and contain variety of information which ultimately falls into big data category. Different statistical methods and database querying approaches have been used in the existing literature for handling big data (Hamel et al., 2005). Traditional data analysis methods depend on assumptions and suitable to solve structured problems. Also those methods are not capable of handling data in huge volume and variety effectively. On the other hand, data mining is a standard process meant of discovering

correlations, patterns, trends or relationships by searching through large data sets stored in data stores, public databases, and data warehouses. Existing literature reveals that, different data mining techniques were successfully used in lifestyle data analysis according to the output try to achieve (Fernández-Villaverde, 2007).

When consider the existing literature, according to a research work carried out in Brazil, early parenthood, child labour, poverty, socio-economic background: education, health, social capital, home violence and employment are identified as factors for school dropout by estimating a logit model (Cardoso et al., 2006). In a similar study which was carried out for Uganda using dimensions such as rural-urban, gender, and age with logistic model analysis found that factors such as rural-urban divide, gender of household head and pupil, age of the household head, household size, academic achievement of mother and father, distance to school, school fees payment, contribution to economy are causes for primary school dropout (Mike et al., 2008). However, both of the studies reflect Inference drawn from statistical hypothesis testing.

The study of Chung and co-workers examine the relationship between chronic diseases and economic outcome (Chung, 2013). This work has studied socioeconomic status on the general health of individuals. Moreover, Chung and others have identified cancer, heart attack, heart disease, lung disease, stroke, arthritis, diabetes, hypertension, psychological problems, asthma, memory loss, learning disability as some of the major chronic diseases which have high impact on family income. However, this study does not focus on patterns on how chronic diseases distributed with the income. There is a similar study on this issue in US (Smith, 1999). This study has focused on impact of health status on economic status statistically. Nevertheless, this has concentrated on health and economic status the near elderly population without working-age population and reveals the patterns specific for those contexts.

Method

Data associated with citizen lifestyle patterns are huge in size and vary in complexity and eventually falls into big data category. When comparing data mining with traditional methods of querying a database, latter require predefine variables. Moreover, data mining is a complicated type of database querying which allows including new and greater number of variables. Also data mining allows examining multiple areas simultaneously. Hence, data mining can be defined as a process of analysing and summarizing data from different perspectives and converting it into useful information (Hariz et al., 2012). Therefore data

mining can be used as a suitable method to analyse citizen life style data which attach with big data problem.

Data mining is a process of analysing big data from different perspectives and summarizing them to useful information by the means of different techniques. The overall process of finding useful information from raw data involves the sequential line up of steps such as developing an understanding of the application domain, creating a target data set based on a smart way of selecting data by focusing on a subset of variables or data samples, data cleaning and pre-processing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, data mining, interpreting mined patterns and consolidating discovered knowledge as in Fig. 1.

Data mining is having two primary goals of being predictive or descriptive. According to the task, different techniques are available in data mining. For predictive tasks, techniques such as classification, regression and deviation detection are used. Meanwhile techniques such as association rules, cluster analysis are used for descriptive tasks. Predictive algorithms determine models or rules to predict the values of variables when given input data. On the other hand descriptive algorithms determine models to summarize the data in some manner. Therefore, selecting the most appropriate mining technique depends on the goal which users going to achieve.

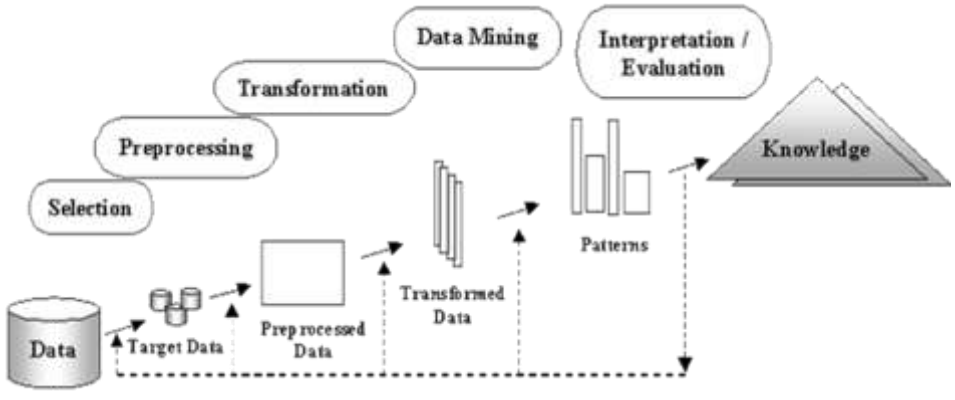

Fig. 1 Steps in Data mining process

Source: http://www.cs.utexas.edu/users/csed/doc_consortium/DC99/wooley-abstract.html

As per the background study main ingredients for the household profiling in Sri Lanka are basic demographic data, income and expenditure for needs and wants of citizens, education details,

health details and housing information. This research of analysing those data using data mining has experimentally done through SPSS and WEKA tools using a 64bit – 2.20 GHz desktop computer. SPSS is proprietary tool used for statistical analysis and WEKA is an open source tool written in Java with machine learning algorithms for data mining tasks such as classification and clustering. Analysing Sri Lankan citizen lifestyle data to find out the issues within lifestyles is identified as the primary research question in this research. Substantively two sub research questions explored within the research. Finding the real factors that contribute to no schooling and school dropout in Sri Lanka using citizen lifestyle data to improve the level of education is one concern. Other concern is analysing huge volume of lifestyle data to determine the relationship between chronic diseases and family income.

In order to build the data model for data mining data selection, preprocessing and transformation are essential steps. Initial attribute selection process for school dropout and no schooling ends up with attributes such as age, household size, district, sector, ethnicity, religion, participation for economic activities, household head's gender, parental education, income, structure of the house and disabilities according to literature. HIES dataset is having missing data due to inconsistency with other recorded data and due to misunderstanding or less value at the time of entry. Consequently, tuples have no recorded value for several attributes such as house number, sex, age. To avoid these incomplete, noisy, and inconsistent data, data pre-processing is done. Then regression is used as the data transformation technique for the data model. Multinomial logistic regression is selected in this case to fine-tuned attributes as dependent variable is nominal with more than two levels with no-schooling and dropouts. The null hypothesis used in the analysis is, there is no difference between the model without independent variables and the model with independent variables. If an attribute having significant level or probability which is less than or equal to the level of significance of 0.05, null hypothesis is rejected and select that attribute for the data model. After extensive pre-processing, the data model is constructed in a manner suitable for two mining tasks: classification and clustering. Then first data mining technique classification is applied to find the patterns in school dropout and no schooling. As the next phase, distribution of chronic diseases according to family income is explored with clustering approach. Data mining algorithms in WEKA tool is used for these analyses. Final results are obtained by comparatively analysing the performance of the well-known mining algorithms in the tool through standard measures.

Findings and Discussion

Factors that contribute to no schooling and dropout according to the statistical logistical regression are in Table 1.

Table 1: Results given by SPSS tool

	Attribute Selected by the Statistical Analysis
	Factors contribute to No Schooling/dropout
No Schooling	Age, District, Religion, Child labour, Disabilities
School Dropout	Age, District, Religion, Child labour, Disabilities, Size of the family, parental education, gender of household head

Then using predictive classification task in data mining we explored the factors that influence no schooling or dropouts. For no schooling and dropouts, separately predictors are selected from HIES dataset. Selected classification techniques for the research are J48 algorithm from decision tree, naïve bayes from Bayesian network classifier and IBk algorithm from K-nearest neighbours. For school dropouts there is only one split in the decision tree which is on the child labour that denotes children who actively participate for the economic activities have high tendency for being dropout from schools. For no schooling religion also matters together with child labour. Results of the classification tasks are in Table 2.

Table 2: Results given by WEKA tool

		Classification Results	
	Algorithm	Correctly Classified Instances (%)	Mean output error
No Schooling	J48	92.93	0.13
	Naive Bayes	92.40	0.09
	IBk	91.78	0.08
School dropout	J48	98.98	0.01
	Naive Bayes	99.02	0.01
	IBk	98.75	0.01

In order to explore the natural distribution of chronic diseases according to family income, for each family main income, main agricultural income, other agricultural income, non-agricultural income and other income attributes in the HIES dataset are considered. Data model

construction process underwent the data pre-processing, data transformation and integration steps. Then, popular K-Means, Expectation-Maximization and Density based clustering approaches have utilized in this analysis. Results of this study pointed out Blood pressure and Diabetics as significant chronic diseases within all Sri Lankans. Moreover, clustering result shows that low income community suffers from many more diseases than high income people. Evaluation of classifier quality is done through standard measurements using in data mining. TP rate, FP rate, Precision, Recall and ROC area is used to determine the predictive accuracy. According to the measurements in Table 3 and 4, Naïve Bayes has the best predictive accuracy in determining school dropout and no schooling.

Table 3: Measures for school dropout dataset

	EVALUATION MEASURES FOR CLASSIFICATION					
	TP RATE	FP RATE	PRECISION	RECALL	F MEASURE	ROC
J48	0.929	0.56	0.935	0.929	0.914	0.684
Naive Bayes	0.924	0.354	0.921	0.924	0.922	0.948
IBk	0.918	0.342	0.917	0.918	0.917	0.81

Table 4: Measures for no schooling dataset

	EVALUATION MEASURES FOR CLASSIFICATION					
	TP RATE	FP RATE	PRECISION	RECALL	F MEASURE	ROC
J48	0.99	0.931	0.985	0.99	0.986	0.53
Naive Bayes	0.99	0.931	0.986	0.99	0.986	0.829
IBk	0.988	0.852	0.984	0.988	0.986	0.85

For the clustering task, times required to build the model and within cluster Sum of Squared Error (SSE) measurement are used to evaluate the data model and technique. Number of clusters where SSE becomes stabilized is chosen as the number of optimum clusters when interpreting final results. Hence we extract patters when data model shows seven clusters

according to Fig 2. The time complexity for K-means, EM and Density Based Clustering are shown in the Table 5. Evaluation shows that K-means is the best algorithm to draw the conclusions as it required minimum time to make cluster in comparison with other algorithms.

Table 5: Runtime measurement for clustering algorithms

	TIME COMPLEXITY FOR CLUSTERING		
	KMEANS	EM	MakeDensityBasedClusterer
Time(seconds)	0.52	8.23	0.54

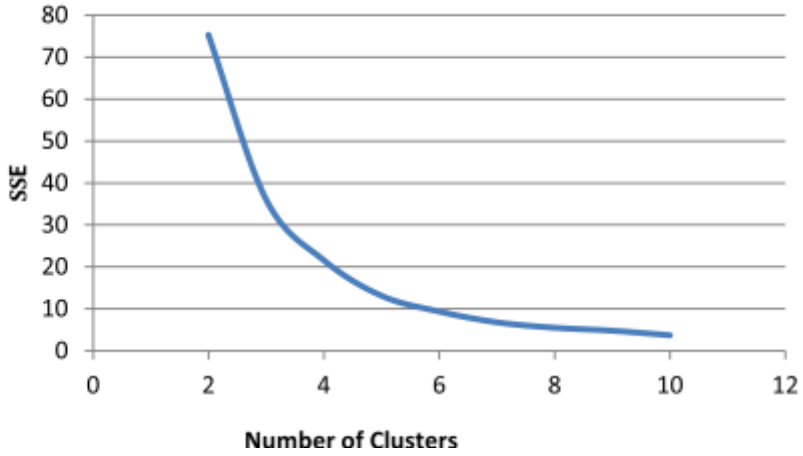

Fig. 2 Scattered graph of SSE vs. Number of clusters

Conclusion

As a developing country Sri Lankan government must set appropriate future development plans to uplift citizen's life quality by carefully analysing socio-economic factors. Furthermore, socio-economic segmentation should be analysed by paying attention to find out hidden pattern behind lifestyle data. Household Income and Expenditure Survey (HIES) dataset is used in this research which cover core fields of Sri Lankan lifestyles to achieve these objectives. When comparing existing solutions given around the world which are most of the time based on statistical approaches, data mining is identified as a novel approach with its ability to analyse big data dynamically and effectively.

Two main cases in lifestyle trends analysis are explored in this research through predictive and descriptive mining. Finding real factors that contribute to no schooling and school dropout in Sri Lanka is the first case addressed through classification mining task. It shows that District, Age, Religion, Child labour and Parental education are major factors related to school dropout and no schooling. In the circumstances where a child is actively participated in income giving

activities school dropout is high. Further, disabilities of the child also directly influence school dropout. School dropouts are high in Colombo, Gampaha and Batticalo districts. Moreover we can see many school dropouts among Buddhist children. The reason could be majority of the sample cases are from Buddhist. No schooling is high in Colombo, Galle and Matara districts according to the analysis. In the second case where we analyse natural grouping of chronic diseases reveals that low income level people suffer from many more diseases than high income people. Furthermore, the clustering task highlights that blood pressure and diabetics as the most common diseases within Sri Lankans. To determine the maximum accuracy of the drawn conclusions different algorithm within the selected techniques are utilized and compare the efficiency before selecting the final the conclusions.

Acknowledgments

We would like to express our gratitude to Department of Census and Statistics, Sri Lanka for their contribution to this research by giving the HIES 2012/2013 data.

References

Cardoso, A. R., & Verner, D. (2006). School drop-out and push-out factors in Brazil: The role of early parenthood, child labor, and poverty.

Central Bank of Sri Lanka. (2013). Central Bank of Sri Lanka Annual Report 2013. Retrieved from: http://www.cbsl.gov.lk//pics_n_docs/10_pub/_docs/efr/annual_report/ar2013/english/content.htm.

Chung, Y., 2013. Chronic Health Conditions and Economic Outcomes.

Department of Census and Statistics. (2013). Household Income and Expenditure Survey -2012/2013 Final Results. Retrieved from: http://www.statistics.gov.lk/HIES/HIES200213FinalBuletin4.pdf

Fernández-Villaverde, J., & Krueger, D. (2007). Consumption over the life cycle: Facts from consumer expenditure survey data. *The Review of Economics and Statistics*, 89(3), 552-565.

Jiang, S., Ferreira Jr, J., & González, M. C. (2012). Analyzing Household Lifestyles, Mobility, and Activity Profiles: A Case Study of Singapore.

Hamel, L., & Hall, T. (2005). A brief tutorial on database queries, data mining, and OLAP. *The Encyclopedia of Data Warehousing and Mining*, 401.

Hariz, M., Adnan, M., Husain, W., & Rashid, N. A. (2012). Data mining for medical systems: a review. *In Proceedings of the international conference on advances in computer and information technology* (pp. 17-22).

Mike, I. O., Nakajjo, A., & Isoke, D. (2008). Socioeconomic determinants of primary school dropout: the logistic model analysis. *African Journal of Economic Review*, 4(1), 217-241.

Smith, J. P. (1999). Healthy bodies and thick wallets: the dual relation between health and economic status. *The journal of economic perspectives: a journal of the American Economic Association*, 13(2), 144.