ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(30), DOI: 10.17485/ijst/2015/v8i30/90251, November 2015* ISSN (Online) : 0974-5645

Mining Children Ever Born Data; Classification Tree Approach

Mahsa Saadati and Arezoo Bagheri*

National Population and Comprehensive Management Institute, Tehran, Iran;
abagheri_000@yahoo.com

Abstract

Classification is an important problem in many fields of science. In this problem the goal is classification of categorical variable as a response based on any input variables as covariates. For years Logistic regression and discriminant analysis were used as statistical methods for classification. Nowadays by developing data base storage and computer programming some new methods such as classification trees are applied to classify data. The aim of this article is introducing CART algorithm as a classification tree rule to analyze demographic data. CARTs generate binary decision tree and have simple interpretations. We use CART to classify children ever born as an important phenomenon in demographic research.

Keywords: CART, Chi-Square Test, Children Ever Born, Classification Tree, Data Mining, Gini Index

1. Introduction

The new generation of advanced computers has capability to save huge data sets. This is a good situation that we can keep any data we want; on the other hand if we can't use information of these data, we won't gain from that. Data mining is a relatively new data analysis method that can encounter with big data sets. Data mining is a collection of various sciences such as computer, artificial intelligence, machine learning and statistics and got its name from the fact that researchers mine among huge data to extract worth information[1]. Data mining is a useful exploratory data analysis process in which the perception and interpretation of pre-determined results of data at hand does not exist. In fact, data mining searches new, valuable and nontrivial information through large volumes of data. Data mining is used both to describe and predict. Explanation the new and nontrivial information based on data is a descriptive aim and providing a model for data is a predictive goal of data mining. In other words, the goal of data mining is prediction, creating a model to classify, estimation or other similar purposes and the main purpose of data mining is description, identification patterns and relationships in data. As mentioned by many researchers, traditional statistical methods cannot replace by data mining, but they can extend by some advanced methods[2].

One of the important issues in data mining, such as statistical analysis, is predictive models. These models based on types of response variable which can be continues or categorical, are divided into regression and classification models. In regression, response mean is predicted by equation that relates response with covariates. When the aim is classification of target variables, probabilities of class membership are estimated in training data set. These probabilities are used to allocate new data that include target variable to different categories[2,3].

Classification tree is a hierarchical and flexible classification method; these characteristics make classification tree as a very attractive and applicable tool for classification. Classification tree splits data into two or more categories based on input covariates, these covariates can be categorical, continuous or any mix of them (e.g., gender, educational level, race, age). The results are often showed in a graph like tree[4,5]. Nowadays classification trees are used in many diverse fields such

* *Author for correspondence*

as social sciences, demography, medicine, business, and biology[6–10]. Although constructing classification tree can sometimes be quite complex and extracted tree may not be so simple, but, the graphical form of it, have simple interpretation even for complex trees. As any predictive model obtaining the most accurate classification model is the goal of a classification tree analysis[11].

This paper outlines an attempt to introduce Classification and Regression Trees (CART) algorithm as a successful classification algorithm in classification tree and apply it to classify Children Ever Born (CEB) data set. It is organized as follows. The following section introduces classification tree induction by CART algorithm, Gini index and chi-square test as measures for determining the best way to classify the data. Moreover, it shows how to prune the tree. Section 3 is about CEB and the application of some statistical methods to analyze it. The results of CART application to analyze CEB are presented in this Section too. Finally, some concluding remarks are presented in Section 4.

2. Classification Tree Induction

Data is splited into smaller divisions called nods by classification tree. Some criteria that is called impurity measure used to divide data, and splitting continue until no more divisions are created. Since classification tree splits data based on one or more than one predictor variables classification tree has multivariate splits on predictor variables[12]. Linear combination splits, can be computed for classification trees when continuous predictors are measured[13]. Classification tree algorithms need some requirements which must be met before applying them; Classification tree algorithms represent supervised learning, and so require preclassified target variables. Data sets divided in two parts; training and learning data. Training data set which is used to extract decision tree should be rich and varied, to provide the algorithm with a healthy cross section of the types of records for which classification may be needed in the future. In classification tree, response (target) variable must be discrete. If target variable isn't discrete, we won't apply classification tree and instead regression tree should be used[14].

2.1 Classification and Regression Trees (CART) Algorithm

Algorithm such as Automatic Interaction Detection (AID), THAID, CHi- squared Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART) can be used for extracting classification trees. AID and THAID are applied for nominal responses, CHAID is used for nominal and ordinal responses and creates multiple splits trees. CART was suggested by[15] and can used for both discrete and continues response variables. CART unlike AID, THAID and CHAID is distribution-free algorithm and more applicable. So in this article we use CART. CART are strictly binary, containing exactly two branches for each classification node. CART recursively divides cases in the training data set into subsets of cases with similar values for the target variable[16]. CART algorithm searches all available variables, all possible splitting values and selects the optimal split for growing the tree; these splitting are done by following criteria[17]:

2.1.1 Gini Index

Gini index is used in tree by binary splitting and for node t and target variable by k categories is defined as:

$$Gini(t) = 1 - \sum_{j=1}^{k} p^2 \left[c = c_{\ j} | \, T = t \right] \quad (1)$$

Where p is the probability that a node t belongs to class C_j and is estimated by $|C_{j,D}|/|D|$ ($|D|$ is the size of subset D). The sum is computed over k categories. The Gini index considers a binary split for each variable[14,16].

2.1.2 Chi-Square Test

The chi-square (χ^2) test for node t and target variable by k categories is defined as:

$$\chi^2(T,X) = \sum_{i=1}^{n} \sum_{j=1}^{k} N\mathrm{x} \frac{\left(p[T_i, c = c_j \,|\, T] - p[T_i \,|\, T] p[c = c_j] \right)^2}{p[T_i] p[c = c_j]} \quad (2)$$

where X is predictor variable, T is subset of nodes and C is number of classification in target variable. When we use (χ^2) as a purity measure its higher value shows that the variability is not occur by chance[14,16].

2.2 Tree Pruning

When a classification tree was constructed, we should consider that the bigger tree is not essentially a better one and this tree may over-fit data; so the solutions must be considered in order to determine the right size for a classification tree. The methods used for this purpose is called pruning. In other words, the optimal tree is optimal

in terms of size and level of classification error. First, estimate the classification error, then using conventional techniques tree pruning and finally, through the pruned tree using the appropriate rule, Tree with the lowest cost and size is selected. Pruning consists of two stages is a classification tree. The first stage is used in ways that would stop the development of excessive classification. These methods are known as pre-pruning. In the second stage useless classification tree branches are removed which is post pruning. Methods used in this phase is to create smaller trees with more accuracy[18,19].

3. Children Ever Born Classification Tree

In demographic research, fertility is one of the important phenomenon. The number of children ever born per woman has important implications for public health, economic, climate, and population structure. It can influence infant, child and maternal mortality, obstetric and child health services, economic growth (or decline), independency burden, labor force participation, and age structure of populations [20]. There is a rich literature about fertility and factors which are had significant effect on it.

In[21] had identified factors which had played major contribution in the enhancement and inhibition of fertility using data of married adolescents in Bangladesh. It was found that early marriage is one of major concern of Bangladesh but it was not sole factor causing adolescent pregnancy. The significant factor effecting adolescent motherhood was education[21]. In[22] had analyzed fertility patterns and its correlates in North East India. Education, religion, occupation, economic status, child mortality, age of women, age at marriage, duration of breastfeeding and use of contraceptives by either spouse were taken as factors which could influence fertility. Status of women had strong influence on fertility[22].

In[23] had analyzed children ever born using binary logistic model by dichotomizing collected from 250 households of slum area of Rajshahi, Bangladesh. Respondents were women of age 15-49. Factors which had contributed significantly to large family were education level of husband and wife, average monthly income and expenditure, ideal number of children, age at marriage and reproductive life span.

In[24] investigated the cause and effect relationship in fertility for Kanyakumari district of India. Initially factor analysis technique was used for grouping of variable into new factors and then multiple regression models were fitted using variables of each group separately. Information related to age of women, age at marriage, religion, type of family, education of husband and wife, work status, income of husband and wife was used.

In[25] had used multiple regression and multilevel analysis to investigate the factors that forced women to high parity in Uttar Pradesh, India. Religion (Islam), women's work status, ever use of contraception and number of child loss had caused increase in fertility while place of residence, women's education, partner's education, type of house, source of lighting had resulted decline in fertility. In Uttar Pradesh it was to be positive estimate the effect of predictors on the parity due to hierarchical structure of data at different levels of survey. In multilevel regression analysis primary sampling unit variable was taken as a regressor along with socio-economic and demographic factors. Both multiple regression and multilevel models produced almost similar results about significance of factors and covariates. The difference was found only between the standard errors. Coefficients of multiple regression analysis had undersized standard errors and consequently resulted in more significant factors than multilevel regression[26].

There is a rich literature about fertility transition in Iran and how expansion in education, reduction in child mortality, urbanization, wide access to family planning services and importance of quality vs. quantity of children have contributed to the recent fertility decline in this country[27–32].

The empirical results of the analysis of[28] represented three groups of determinants influence fertility behavior of Iranian households. The first group consists of economic factors either at micro or macro levels. Second, distribution of intra-household bargaining power has a strong influence on fertility in Iran. Finally, although there was no difference between the number of children in urban and rural areas, the findings yield a support for the role of other demographic determinants such as literacy, social norms of household size, and religion on fertility behavior of Iranian families[27].

Since there isn't any study which is used classification tree to analyze CEB in Iran, also the benefit of this model for classification in this article we use CART to classify CEB.

3.1 Result

In this study, Children Ever Born (CEB) in survey of "Study Marriage and Fertility Attitudes of Married 15-49 Year-Old Women in Semnan, Iran; 2012"[33], are classified by classification tree approach. The data in this survey is achieved by a cross-sectional survey in Semnan province which has been collected by a structural questionnaire. Semnan is a province that is taking efficient steps to development and modernization. Nowadays, it is considered as one of the developed province in Iran. In this province, changes in fertility attitudes and beliefs expected to be affected by the modernization, industrialization and urbanization[33]. 405 samples from 2 cities and 6 villages of Semnan province, among 8 cities and 589 villages, were selected. This sample includes of 15-49 year-old women who has married in private settled household. CEB, age at first marriage, marriage type, education levels, job status, birth Place and birth cohort were collected in this research.

Table 1 shows frequency and percentage of categorical variables. CEB of 45 percentages of women is 2 children. Almost equal percentage of women exists in all three birth cohorts. 80, 67.1 and 77.3 percentages of women were unemployed, diploma and above and born in urban area respectively. Nearly 60 percentages of women had non-familial marriage type.

Table 1. Frequency table of study variables

Percent	Frequency	Variable	
		Value	Name
11.4	46	0	Children Ever Born (Response Variable)
21.2	86	1	
45.2	183	2	
22.2	90	3+	
32.3	131	1960	Birth Cohort
35.8	145	1970	
31.9	129	1980	
19.8	80	Employed	Job Status
80.2	325	Unemployed	
67.1	134	Diploma and above	Educational Levels
33.0	271	Less than Diploma	
59.3	240	Non-Familial	Marriage Type
40.7	165	Familial	
77.3	313	Urban	Birth Place
22.7	92	Rural	
405		**Total**	

Mean and standard deviation of age at first marriage variable are 20.75 and 3.15, respectively.

3.2 CART Classification Tree for Classification of CEB by Gini and chi-square Indices with Estimated Prior Probabilities

Figure 1, 2 presents classification tree of CEB according to predicted variables of age at first marriage, marriage type, educational levels, job status, birth place, and birth cohort by Gini index and Chi-square test with estimated prior probabilities respectively. All of the predicted variables are entered in this classification tree as nodes. Birth cohort has been placed in the root of the classification tree as the most influenced variable on classifying CEB.

Table 2 presents the misclassification matrix of Model 1 and 2 which indicates the accuracy of two classification models. The shaded cells in Table 2 signify correct classification or accuracy of the classification trees on Figure 1 and 2. The accuracy of the classification tree can be calculated as Equation (3) and (4) for Model 1 and 2 respectively:

$$\text{Accuracy}_{\text{model(1)}} = \frac{18+49+143+55}{405} = 0.65 \quad (3)$$

$$\text{Accuracy}_{\text{model(2)}} = \frac{18+49+115+76}{405} = 0.64 \quad (4)$$

0.65 of classification accuracy Model 1 means that CEB of 65 percentages of women has been classified correctly. This value indicates that misclassification is equal to 35 percent. As Equation (3) and (4) present, the accuracy of two models are approximately equal. But Model 1 is more complex than model 2; Model 1 contains 11 nodes and 12 leaves compare with Model 2 with 8 nodes and 9 leaves. So we recommend using Model 2 instead of Model 1.

The following rules can be extracted from the classification tree in Figure 2:

- CEB of women in the first birth cohort (1960 decade) was 3 and more children without affects of any other predictor.
- CEB of women in the second birth cohort (1970 decade) whose age at first marriage was low (≤15.5) was 3 and more children. Birth place for women in this cohort whose age at first marriage was high (>15.5) with both educational levels didn't play any specific rules in classifying CEB. Their CEB were 2 children either they were living in urban or rural areas.

Table 2. Misclassification matrix for classification

Total	Model (2) Predicted Category				Total	Model (1) Predicted Category					Observed Category
	3^+	2	1	0		3^+	2	1	0		
46	6	10	12	18	46	5	11	12	18	Numbers	0
	4.14	6.41	17.65	50.00		5.68	5.16	17.65	50.00	Column Percentage	
	13.04	21.74	26.09	39.13		10.87	23.91	26.09	39.13	Row Percentage	
11.36	1.48	2.47	2.96	4.44	11.36	1.25	2.72	2.96	4.44	Total Percentage	
86	8	17	49	12	86	24	1	49	12	Numbers	1
	5.52	10.90	72.06	33.33		1.14	11.27	72.06	33.33	Column Percentage	
	9.30	19.77	56.98	13.95		1.16	27.91	56.98	13.95	Row Percentage	
21.33	1.98	4.20	12.10	2.96	23.21	0.25	5.93	12.10	2.96	Total Percentage	
183	55	115	7	6	183	27	143	7	6	Numbers	2
	37.93	73.72	10.29	16.67		30.68	67.14	10.29	16.67	Column Percentage	
	30.05	62.84	3.83	3.28		14.75	78.14	3.83	3.28	Row Percentage	
45.19	13.58	28.40	1.73	1.48	45.19	6.67	35.31	1.73	1.48	Total percentage	
90	76	14	0	0	90	55	35	0	0	Numbers	3^+
	52.41	8.97	0	0		62.50	16.43	0	0	Column Percentage	
	84.44	15.56	0	0		61.11	38.89	0	0	Row Percentage	
22.22	18.77	3.46	0	0	22.22	13.58	8.64	0	0	Total Percentage	
405	145	156	68	36	405	88	213	68	36		Total
100	35.80	38.52	16.79	8.89	100	21.73	52.59	16.79	8.89		Total Percentage

- CEB of unemployed women in the third birth cohort (1980 decade) whose with non-familial marriage and familial marriage were 1 and 0 child, respectively.
- CEB of employed women in the third birth cohort (1980 decade) whose age at first marriage were low (≤24.5) and high (>24.5) were 2 and 1 children, respectively.

Risk and standard error of classification tree for training and learning data and estimated prior probability have been shown in Table 3 and 4, respectively. According to the results of Table 3, these values are almost equal which indicates the validity of classification model proposed by classification tree in Figure 2.

Table 3. Risk and standard error of classification tree for training and learning data

	risk	Standard error
Learning set	0.363	0.024
***k*-fold cross validity of training set**	0.339	0.025

Table 4. Estimated prior probability of classification tree

Category	Prior probability	Number
0	0.114	46
1	0.212	86
2	0.451	183
3^+	0.222	90

4. Conclusion

Classification trees recursively partitioning predictor variables to separate areas and allocate data to classes. The recursive partitioning lead to a fixed piece model on the predictive variable space. For partition each node, all partitions validate for each predictors. Variable and corresponding partitioning points select in such a way that the best separation between two nodes is obtained.

This process continues recursively until each node contains a limited number of cases. After making a big tree, the rules for pruning and adjusting the size of the tree is used. Classification and Regression Trees (CARTs) are useful in generating binary classification trees by splitting the subsets of the dataset using all predictor variables to create two child nodes[3]. CART algorithm is an important method that uses classification and regression tree analysis of large data sets. CART has so many advantages such as:

- CART doesn't have any distributional assumption on covariates and response variables.
- Covariates can be mixed of both categorical and continuous variables.
- CART can deal with missing data by some methods, so no case is deleted from the analysis because of missed information.
- CART doesn't affected by outliers.
- It can consider interaction effect between variables[3,11,16].

Classification tree have been used for classification in many fields, such as demography, medicine, manufacturing and production, financial analysis, astronomy, and molecular biology[11].

Classification trees are different from discriminant analysis and logistic regression which are traditional statistical methods and used to classify data. Both discriminant analysis and logistic regression need some assumptions that without them the validity of results is not met. For example in discriminant analysis all covariates must be continuous and have normal distributions that are not confirmed in many applied studies. Validation of logistic regression depends on enough cases in each target variable categories and when the number of covariates increases, the full models that contains interactions gets more complex. In classification tree any assumptions about distributions of response and covariate variables does not need. Also we can consider interactions between variables without complexity[13,14].

In this study, Children Ever Born (CEB) in survey of "Study Marriage and Fertility Attitudes of Married 15-49 Year-Old Women in Semnan, Iran; 2012" are classified by classification tree approach. The following results have extracted from CBR classification tree:

- CEB of women in the first birth cohort (1960 decade) was 3 and more children without affects of any other predictors. Although CEB of women in the second birth cohort (1970 decade), whose married in low and high age, was 3 and more and 2 children respectively.
- Marriage type has not affected on CEB in first and second birth cohort which are 2 or 3 children and more which only depend on age of their marriage.
- CEB of women in the third birth cohort (1980 decade) was affected by type of marriage while women whose type of marriage is non-familial, have 1 child, women with familial marriage were childless.
- 1 or 2 were CEB of employed women in third cohort whose age of marriage were low and high respectively. Employed women by low age of marriage comparing to unemployed women had higher CEB.

5. Acknowledgment

This article is extracted from a survey under the title of "Mining Demographic Data by Decision Tree" which is supported by National Population Studies and Comprehensive Management Institute in 2014 by the registered number of 20/15283.

6. References

1. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. 3rd ed. Potomac, MD; 1999.
2. Berry MJA, Linoff GS. Data mining techniques: For marketing, sales, and customer support. New York: John Wiley and Sons; 1997.
3. Larose DT. Data mining methods and models. John Wiley and Sons Inc Publication; 2006.
4. Berry MJA, Linoff GS. Mastering data mining: The art and science of customer relationship management. 2nd ed. New York: John Wiley and Sons; 1999.
5. Linoff GS, Berry MJA. Mining the web: Transforming customer data into customer value. New York: John Wiley and Sons; 2002.
6. Kim H, Loh WY. Classification trees with bivariate linear discriminate node models. Journal of Computational and Graphical Statistics. 2003; 12:512–30.
7. Nanni **L**, Lumini **A**. Manna **C**. A data mining approach for predicting the pregnancy rate in human assisted reproduction. Advanced Computational Intelligence Paradigms in Healthcare. 2011; 326:97–111.
8. Gams M, Krivec J. Demographic analysis of fertility using data mining tools. Informatica. 2008; 32:147–56.
9. Mello F, Bastos L, Soares S, et al. Predicting smear negative pulmorary tuberculosis with classification trees and logistic regression: a cross-sectional study. PMC Public Health. 2006; 23:6–43.
10. Gould M, Affiliation: Kaiser Permanente Southern California, Pasadena, California, United States of America
11. Khazeni N. Demographic and clinical predictors of mor-

tality from highly pathogenic avian influenza A (H5N1) virus infection: CART analysis of international cases; 2014. Available from: www.plosone
12. Olson DL, Delen D. Advanced data mining techniques. Berlin Heidelberg: Springer-Verlag; 2008.
13. Buntine W, Niblett T. A further comparison of splitting rules for decision-tree induction. Machine Learning. 1992; 8:75–85.
14. Hand DJ, Mannila H, Smyth P. Principles of data mining. Cambridge, USA: MIT Press; 2001.
15. Jiawei H, Jay B. Data mining: Concepts and techniques; 2nd ed. University of Illinois at Urbana-Champaign; 2006. p. 285–484.
16. Breiman L, Jerome F, Richard O, Charler S. Classification and regression trees. Chapman and Hall; 1984.
17. Bringman B, Zimmermann. Tree decision tree for tree structured data. Springer; 2005. p. 46–58.
18. Kennedy RL, Lee Y, Van Roy B, Reed CD, Lippman RP. Solving data mining problems through pattern recognition. Upper Saddle River, NJ: Pearson Education; 1995.
19. Frank. Data mining: Practical machine learning tools and techniques; 2nd ed. 2005.
20. Gilbert R. CHAID and Earlier Supervised Tree Methods; 2010. Available from: http://www.unige.ch/ses/metri/
21. Cleland JG. Trends in human fertility. In Heggenhougen HK, editor. International Encyclopedia of Public Health. Oxford: Academic Press; 2008. p. 364–71.
22. Abedin S, Rahman JAM. On the dynamics of high-risk fertility in Bangladesh. International Journal of Human Science. 2010; 9.
23. Dey S, Goswami S. Fertility pattern and its correlates in North East India. Journal of Human Ecology. 2009; 26(2):145–52.
24. Rahman M. Predicting the number of children ever born using logistic regression model. Biometrics and Biostatistics International Journal. 2015; 2(4).
25. Senthamarai KK, Nagarajan V. Factor and multiple regression analysis for human fertility in kanyakumari district. Anthropologist. 2008; 10(3):211-4.
26. Dwivedi SN, Rajaram S. Some factors associated with number of children ever born in Uttar Pradesh: A comparative results under multiple regression analysis and multilevel analysis. Indian Journal of Community Medicine. 2004; 29(2):72–6.
27. Logubayom IA, Luguterah A. Survival analysis of time to first birth after marriage. Research on Humanities and Social Sciences. 2013; 3(12).
28. Abbasi-Shavazi MJ, McDonald P, Hosseini-Chavoshi M. The fertility transition in Iran: Revolution and reproduction. 2st ed. Canbera: Springer. National University Canbera; 2009. p. 48–50.
29. Abbasi-Shavazi MJ, Torabi F. Women's education and fertility in Islamic countries population dynamics in muslim countries. Springer; 2012. p. 43–62.
30. Aghajanian A. A new direction in population policy and family planning in the Islamic Republic of Iran. Asia-Pacific Population Journal/United Nations. 1995; 10(1):3.
31. Aghajanian A, Mehryar AH. Fertility transition in the Islamic Republic of Iran: 1976-1996. Asia-Pacific Population Journal/United Nations. 1999; 14(1):21.
32. Salehi-Isfahani D, Abbasi-Shavazi MJ, Hosseini-Chavoshi M. Family planning and fertility decline in rural Iran: The impact of rural health clinics. Health Economics. 2010; 19(S1):159–180.
33. Torabi F. Marriage postponement and fertility decline in Iran. London School of Hygiene and Tropical Medicine (University of London); 2011.
34. Razeghi H. Marriage and fertility behavior at least once married women, 15-49 years old in 1391 Smnan-Iran. Research report; 2013.