CHAID UNDER HEALTH CARE SYSTEM

N.Hema [1],Jha[2]

[1] Assistant Professor, Dept. Of Computer Science,
Vivekanandha College Of Arts And Sciences For Women, Elayampalayam, Tiruchengodu(Tk),
Namakkal (Dt.) Tamilnadu, India.

[2] Professor,Department of Computer Science,
M.L.S.M College
Darbanga, Patna. India

Abstract

Data mining application in healthcare today is great, because healthcare sector is rich with information, and data mining is becoming a necessity. Healthcare organizations produce and collect large volumes of information on daily basis. Use of information technologies allows automatization of processes for extraction of data that help to get interesting knowledge and regularities, which means the elimination of manual tasks and easier extraction of data directly from electronic records, transferring onto secure electronic system of medical records which will save lives and reduce the cost of the healthcare services, as well and early discovery of contagious diseases with the advanced collection of data. Healthcare data mining provides countless possibilities for hidden pattern investigation from data sets. These patterns can be used by physicians to determine diagnoses, prognoses and treatments for patients in healthcare organizations. In the application of data mining in explaining women's choice of contraceptive methods, we used CHAID algorithm. This paper describes the CHAID under health care system.

Keywords: Data mining, Healthcare

,

.

1. INTRODUCTION

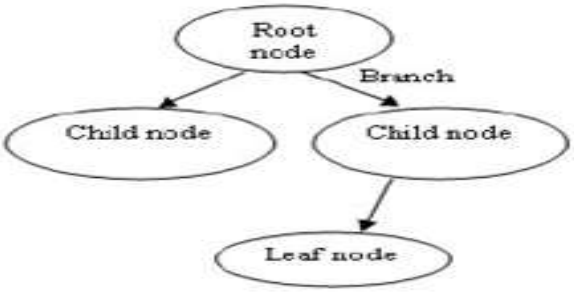
Healthcare organizations today are capable of generating and collecting large amounts of data. This increase in volume of data requires automatic way for these data to be extracted when needed. With the use of data mining techniques it is possible to extract interesting and useful knowledge and regularities. Knowledge acquired in this manner, can be used in appropriate area to improve work efficiency and enhance quality of decision making process. In this manner we implement the CHAID algorithm.

1.1 Decision tree

Decision tree is a graphical representation of the relations that exist between the data in the database. It is used for data classification. The result is displayed as a tree, hence the name of this technique. Decision trees are mainly used in the classification and prediction. It is a simple and a powerful way of representing knowledge. The models obtained from the decision tree are represented as a tree structure. The instances are

ISSN : 0975-8283

classified by sorting them down the tree from the root node to some leaf node [21]. The nodes are branching based on if-then condition. Tree view is a clear and easy to understand, a decision tree



1.2 CHAID

CHAID is the oldest algorithm of the mentioned tree, and it was firstly published by Hartigan in 1975. It is derived from the AID (Automatic Interaction Detection) with a purpose to detect statistical relations between variables (which is done by building a decision tree) and it has been used for classification (8).

CHAID is the oldest algorithm of the mentioned tree, and it was firstly published by Hartigan in 1975. It is derived from the AID (Automatic Interaction Detection) with a purpose to detect statistical relations between variables (which is done by building a decision tree) and it has been used for classification.

As mentioned above, CHAID is different from other algorithms because it follows prepruning method, that is; it tries to stop the branching before the over fitting (inability to generalize the results based on training set to other data) occurs. Another difference is that CHAID works only with categorical variables so it is necessary to break the ranges in the continued ones, or replace the key to CHAID lies in its first two letters that mark the test of significance that is X^2 (chi square test) . CHAID uses X^2 test to make a decision about merging the fields that do not create statistically significant differences in the values of a target field. After that, it splits every group with three or more fields with binary splits, and if some of these splits generate a statistically significant difference in outcomes, CHAID retains that split. Next, X^2 test is being used again, and the field that generates the groups that differentiate the most (according to that test) is being chosen as a splitter for that node. The tree is growing until there are no more splits that lead to statistically significant differences in classification. Precise level of significance determines the size of a tree and its value as a classifier (8).

1.3 Usage of data mining in health care management

Table 1 contains different categories of data mining applications in health care. The biggest number of papers is on drug development and research (12%), data modeling for health care applications – e.g. nursing (11%), and executive information systems for health care (10%). Pub- lic health informatics applications are described in 9% of articles, and e-governance structures in health care is subject of 8% of articles. Forecast- ing treatment costs and demand of resources is presented in 7% of articles. Other articles contain less than 5% of the sample. However, there is approximately one forth of the articles that could not be classified in homogenous category.

Table 1. Categories of data mining applications in health care

Category	No of Articals (%)
Drug Development and Research	27(12)
Data Modeling for healthcare Applicatons .Eg-Nursing	24 (11)
Executive Information System	22 (10)

ISSN : 0975-8283

for health care	
E-governance in health care	18 (8)
Public health informatics	20 (9)
Forecasting treatment costs and Demand of resources	46 (21)
Demonstration of data mining software Application in healthcare	11(5)
Other	53(24)
Total	**221(100%)**

1.4 Decision tree on the choice of contraceptive method

The database consists of 1,473 cases. Variables of interest are: women's age, education (1=low, 2, 3, 4=high), husband's education (1=low, 2, 3, 4=high), number of children ever born, religion (0=non Islamic, 1=Islamic), employment (0=not employed, 1=employed), hus- band's profession (1=low, 2, 3, 4=high), index of living standard (1=low, 2, 3, 4=high), exposure to media (0=good, 1=not good), method of contraception used (1=no usage, 2=long term method,3=short term method).

Fore In the example of „choice of contraceptive method", CHAID did the first split on „husband's profession", which was obviously the most important. In ID=1, CHAID started with 1,473 cases where non usage of contraceptive methods prevailed, and it split the data into two significant groups. Women whose husbands work on a hierarchically low places were grouped in ID=2 group (436 instances) and other 1,037 cases were placed in ID=3 group. In both groups, non us- age of contraception prevails, but it is important that in ID=2 group, the long term method usage is significantly higher than in ID=3 group. Fur- ther on, ID=2 group have been broken down by number of children, on ID=4 group (number of children is less or equal to 2) where there are 187 instances where non usage prevails, and ID=5 group (number of children greater than 2) that has249 women that mostly use a long term method of contraception. ID=5 group have been broken down by woman's' education (ID=17 and ID=18) where the left side of the tree ends. ID=17 group represents 82 women with high education where non usage of contraception prevails, while ID=18 group represents 157 women that mostly use long term method.

Hierarchically higher ranks of employment of husbands (ID=3) have been broken down by „ed- ucation" into „very low" and „low" (ID=6) that has 442 instances where non usage prevails. ID=6 is further divided on ID=9 and ID=10 (according to „media exposure") that has been ranked as "high"(ID=9, sample of 363 women) where „non usage"prevails and „high" (ID=10, sample of 79 women) where „non usage" also prevails.

ID=7 has 329 women that mostly do not use contraception and it has been split by „woman's age" into two groups ID=11 and ID=12. ID=11 has 218 cases that are younger than 32 years, and usage of short term methods prevails in this group. Furthermore, it has been broken down by „working status of a woman" on ID=13 (YES) and ID=14 (NO). There are 50 working women in ID=13 where „non usage"prevails. In ID=14, there are 168 women currently unemployed, and the short term methods are preferred here. ID=14 has been split by „husbands' profession" on ID=15 and ID=16. ID=15 includes women whose husbands work on a medium level of hierarchy, and there are 167 cases where a short term method prevails. It is interesting that if a hierarchical level grows on high, there is one woman in ID=16 that uses a long term method. Once again we have the „husbands' profession" as a very im- portant variable. If we go up the tree on ID=12, we will notice a group that has 111 women, where „non usage" prevails. Furthermore, ID=8 counts 266 women who mostly use short term methods.

ISSN : 0975-8283

2. MODEL EVALUVATION

In order to establish how well the model is working, we created a „Disagreement Table for a Predicted Variable“ that shows a percentage of prediction deviation (Table 2) and “Goodness of Fit “ tables that show how many cases are classified correctly/incorrectly (Table 3, Table 4).

Table 2. Frequency table for predicted variable “Method of contraception” by CHAID

Observed	Predicted	No
Non usage	Non usage	109
Non usage	Long term	9
Non usage	Short term	57
Long term	Non usage	26
Long term	Long term	35
Long term	Short term	24
Short term	Non usage	51
Short term	Long term	20
Short term	Short term	50

Table 3. Disagreement table for predicted variable “Method of contraception” by CHAID

Method	Deviation percent
Non usage	41.39785
Long term	45.3125
Short term	61.83206

Table 4. Disagreement table for observed variable “Method of contraception” by CHAID

Method	Predication by CHAID
Non usage	37.71429
Long term	58.82353
Short term	58.67769

3. DISCUSSION

With assistance of the decision tree we have managed to get some valuable information that we could not have noticed at first glance with use of classic charts. One of these data is that the key variable is „a hierarchical position of a husband on working place“. To some point it is understandable since the data consists mostly of women that belong to Islamic religion, where husband’s domination is common. That is why we are mostly concerned with the population of women whose’ husbands work on a hierarchically „low“, „medium“ or „higher“ position (excluding „high“). Recent Australian reports confirm that even developed countries dis- cuss issues of a women’s choice of contraception or even of willingly having sex. Another thing of importance is the “role power” of motherhood, as it provides a certain status of maturity (giving sense of control over a mother’s and child’s life to a woman, as well as over her income) especially in depressed areas (11). If we look at ID=10 in Figure 2, we can argue that the dominance of husbands prevails in cases of woman’s low education and her low exposure to the media (which is of high impor- tance concerning non usage when ID=6, ID=7 and ID=8 are compared).

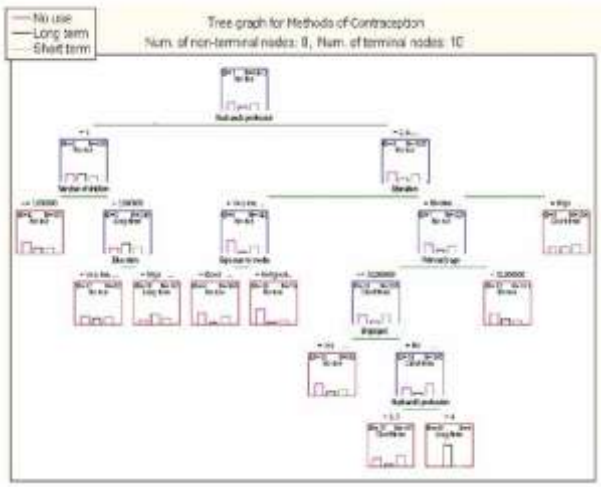


Figure 2. CHAID decision tree for the choice of contraceptive method

Women’s age “also has an important role in this case, since, until the age of 32, women

ISSN : 0975-8283

mostly use short term methods and after that non usage "prevails. Further on, CHAID has divided a population of women younger than 32 according to their status of employment", where working women mostly do not use contraception and unemployed women mostly use short term methods.

According to these results, we think that the health care management could raise the awareness of women conducting a campaign for contraceptive products. A marketing strategy should model developed. Although, we have searched many scientific databases, there are many com- mercial applications of data mining in health care that are not included in this review. Therefore, one has to be careful in interpreting the results of this analysis. However, it is by no means an illustration of what is going on in the field of data mining applications in health care. In addition, decision trees can be quite complex, which makes the interpretation of the rules rather complicateded and problematical.

REFERENCES

1. International Conference on Population and deveopment, Cairo, Egypt, 1994). [http://www. un.org/popin/icpd2.htm]
2. amberger D, Šmuc T, Marić I. DMS - poslužiteljza analizupodataka, Institut Ruđer Boškovic,2001. /tutorial/hr_tut_glosary.php]
3. Ross J, Hardee K, Mumford E, Eid S. Contracep- tive method choice in developing countries. IntFam Plan Perspect 2002; 28:32-40.
4. H. Witten and F. Eibe, Data Mining: Practical Machine Learning Tools and Techniques, vol. 2, Diane Cerra Publishers, 2005.
5. Lim TS. Contraceptive method choice, 1987. [http://www.ics.uci.edu/~mlearn/MLSumma ry. html]
6. Seiber E, Bertrand J, Sullivan T. Changes in con- traceptive method mix in developing countries. Int FamPlan Perspect 2007; 33: 117–123.
7. .A. Walter, "Data Mining Industry: Emerging Trends and New Opportunities", Massachusetts Institute of Technology, pp. 13-15, 2000.
8. Kantardzic, Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2003.
9. Hamilton HJ. Overview of decision trees. CS 831, 2003.[http://www2.cs.uregina.ca/~dbd/cs83 1/ notes/ml/dtrees/4_dtrees1.html]
10. University of Maryland. Public health informatics, University of Maryland, 2000. [http://www.phi.umd.edu/what/]

ISSN : 0975-8283