Expert Systems with Applications 38 (2011) 1400–1404

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Mining the optimal clustering of people's characteristics of health care choices

Chieh-Yu Liu [a,b,*], Jih-Shin Liu [c]

[a] Department of Nursing, National Taipei College of Nursing, 365, Min-der Rd., Bei-Tou District, Taipei City 112, Taiwan, ROC
[b] Department of Health Care Management, National Taipei College of Nursing, 365, Min-der Rd., Bei-Tou District, Taipei City 112, Taiwan, ROC
[c] Division of Biostatistics and Bioinformatics, National Health Research Institutes (NHRI), 35, Keyan Rd., Zhunan Town, Mioli County 350, Taiwan, ROC

ARTICLE INFO

Keywords:
Health care
k-Means cluster analysis
v-Fold cross-validation
Multiple Correspondence Analysis

ABSTRACT

In Asian countries, there has been a multi-choice healthcare environment for many years. In Taiwan, people's multiple health care seeking behavior has resulted in much heavier financial burden of National Health Insurance Program (NHIP) in recent years: investigating the characteristics of people who use multiple health care resources has gained increasing importance for health authorities. In this study, we investigated the socioeconomic and demographic characteristics which underlined people's choice of health care by using a population representative database. A novel methodology which incorporated *k*-means cluster analysis with *v*-fold cross-validation into Multiple Correspondence Analysis (MCA) is proposed. This novel methodology can help us to find the optimal attribute clustering of multiple health care utilization. By using this methodology, researchers not only can avoid the ambiguities of identifying clusters resulted from the traditional hierarchical cluster analysis (HCA), but also can provide more solid and evidence-based analysis for health policy making.

1. Introduction

In Asian countries, there has been a multi-choice environment of health care for a long time. Although the prevalence of or why people used complementary and alternative medicine (CAM) has been extensively studied in western countries in recent decades (Astin, 1998; Barnes, Powell-Griner, McFann, & Nahin, 2004; Eisenberg et al., 1993; Honda & Jacobson, 2005). In Asian countries, this has received much less attention. In Taiwan, the financial burden of National Health Insurance Program (NHIP) became much heavier in recent years so investigating the characteristics of people who consumed or utilized more NHI resources has become a very important health policy-making topic and investigating the socioeconomic and demographic factors underlying people's choice behavior among different health care gains increasing importance.

Liu, Chen, Weng, and Liu (2006) ever used Multiple Correspondence Analysis (MCA) and traditional hierarchical cluster analysis (HCA) to identify the attribute clusters of people's characteristic among different health care choices. Although such methodology has been popularized by social science researchers (Guinot et al., 2001; Milani, Cortinovis, Rainisio, Fognini, & Marubini, 1983; Minoura, 1987; Roux et al., 1995), however, this methodology could result in some drawbacks: (1) as the number of study subjects increases, it is more difficult to identify clusters from a tree diagram derived from HCA; and (2) researchers identify the resulting clusters from traditional HCA according to their subjective interpretation. These drawbacks cast doubts on the validity of this method.

Correspondence analysis (CA) has been widely applied for analyzing questionnaire data, especially for those with multiple choice questions (Greenacre, 1984). CA can be classified into Simple Correspondence Analysis (SCA) and Multiple Correspondence Analysis (MCA). SCA is a descriptive or exploratory technique designed to analyze two-way tables containing some measurements of correspondence between the rows and columns. The SCA results provide information that is similar in nature to those derived from Factor Analysis techniques, which allows one to explore the structure of categorical variables included in the table or space. MCA is the generalized extension of SCA to deal with more than two variables. Interpretation of MCA's results is similar to those for SCA; but the increased complexity of the data requires an added dimension to any analytical process. The purpose of this paper is to propose a novel methodology that incorporates *k*-means cluster analysis with *v*-fold cross-validation into MCA for determining the optimal clustering of people's characteristics among different health care choices. This algorithm consists of three parts: (1) obtaining a set of coordinate values summarizing respondents' demographic and socioeconomic characteristics associated with their health care choices, allowing associations between demographic and socioeconomic characteristics and health care choices to be displayed graphically; (2) using *k*-means cluster method (Everitt, 1993) to

* Corresponding author at: Department of Nursing, National Taipei College of Nursing, 365, Min-der Rd., Bei-Tou District, Taipei City 112, Taiwan, ROC. Tel.: +886 2 2822 7101x3312; fax: +886 2 2821 3233.
E-mail address: chiehyu@ntcn.edu.tw (C.-Y. Liu).

0957-4174/$ - see front matter Crown Copyright © 2010 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.07.044

cluster respondents' demographic and socioeconomic characteristics and their health care choices, however, the optimal number of clusters is unknown; and (3) applying the v-fold cross-validation method (Efron & Tibshirani, 1993) to estimate the optimal number of clusters and help to determine the optimal clustering.

2. Data and methods

2.1. Data

This study used the data from the National Health Interview Survey conducted in 2001 which is representative of the general Taiwanese population. Taiwan had launched a pilot NHIS in 1994, but after that, such nationwide health survey has not been conducted until 2001. The purpose of this survey was to understand the general health status of the resident civilian non-institutionalized population in Taiwan. And since 2001, it had become a periodical survey every 4 years. The launch of NHIS was a new milestone for Taiwan public health administration. It mainly surveyed four aspects: health status, health care resources utilization, health behavior and quality of life in Taiwan. 6364 households (23,473 persons) were sampled from the whole Taiwan area. The demographic characteristics of the samples were consistent with the population in Taiwan. The response rate was 91.1% for households (5798 households) and 94.2% for individuals (22,112 individuals). Details of this survey have been published by Shih et al. (2003). In this study, we used the subset of adults aged over 20 years old and who provided feasible answers to the questions on the questionnaire. A total of 15,413 adult respondents were eligible.

We mainly studied the choices of health care when one adult respondent has the minor illness, such as headache, abdominal pain, diarrhea or common cold. The possible answers were:

(1) Consult a Conventional western medicine (CWM) physician.
(2) Consult a Traditional Chinese medicine (TCM) physician.
(3) Go to a pharmacy for Over-The-Counter (OTC) medication.
(4) Seek for Folk medicine (such as foot massage, shave cholera, or ask for a shaman or ask for a spell).
(5) Ignore it.
(6) Others, such as ––––––––.

Up to two of these five mostly used choices were allowed. So this is a multiple-choice question design.

2.2. Methods

The analysis steps and the novel algorithm proposed in this paper are illustrated below:

Step 1. The descriptive statistics of respondents' demographic and socioeconomic characteristics were tabulated.
Step 2. The MCA was applied giving a set of coordinate values summarizing respondents' demographic socioeconomic and health care choice characteristics. This allowed the associations to be displayed graphically by plotting the scree plot of eigenvalues derived from the MCA. This allows determination of the number of dimensions of MCA to be visualized.
Step 3. k-Means cluster analysis with v-fold cross-validation was applied to analyze the MCA coordinates of the demographic and socioeconomic characteristics associated with health care choices, the algorithm is as follows:

(1) Setting the number of cluster for k-means cluster algorithm: started from $k = 2$ to n, $n \leqq 10$.
(2) Setting the selection of cluster centers: the initial k cluster centers were selected at random.
(3) Assigning all subjects to their nearest cluster centers. The squared Euclidean distance is used as the distance measurement in these calculations.
(4) v-Fold cross-validation was then applied to estimate the optimal number of attribute cluster and the optimal clustering:
  (a) Dividing the sample ($\boldsymbol{S}$) into v folds (denoted S_i, $i = 1, \ldots, v$), in this study, we used the $v = 5$.
  (b) For $i = 1$ to v, take S_i as the testing set, $\{\boldsymbol{S}\}$ S_i as the training sets.
  (c) Computing the mean squared Euclidean distances within each cluster of training sets, letting these be the new cluster centers and replacing the cluster centers of the previous step.
  (d) Computing mean distances (squared Euclidean distance) of each point of testing set to the new cluster centers derived from the training sets.
(5) Repeat from (1).
(6) If $k = j$ who has the minimum mean distances of each point of testing set, j would be the optimum number of clustering.
(7) Stopping rule (Hsu & Ray, 1999): If $|\overline{D}_{j+1} - \overline{D}_j| < 0.01$, then stop dividing and clustering.

The algorithm we proposed in this paper was implemented by using STATISTICA Data Miner ver 6.0 (Tulsa, OK, USA).

3. Results

Demographic and socioeconomic characteristics of the 15,413 respondents are presented in Table 1. 50.70% were females and

Table 1
Socio-demographic characteristics of the whole study sample ($N = 15{,}413$).

Characteristics	No.	(%)
Sex		
Female	7815	50.70
Male	7598	49.30
Age (yr)		
20–29	3552	23.05
30–39	3482	22.59
40–49	3373	21.88
50–59	2105	13.66
≧60	2901	18.82
Education		
<High School	6980	45.31
High School	4526	29.39
≧College	3896	25.30
Missing	11	
Monthly Income[a]		
<NTD$20,000	8010	52.28
NTD$20,000–39,999	4316	28.17
NTD$40,000–59,999	1965	12.83
NTD$60,000–79,999	653	4.26
≧NTD$80,000	376	2.45
Missing	93	
Religion		
Buddhism, Dao and Other Chinese Tradition Religions	13687	89.04
Catholic and Christian	890	5.79
Other religions	38	0.003
Atheists	756	4.92
Missing	42	
Occupational status		
Hired by government	1085	7.08
Hired by non-government	5360	34.96
Self hired	2645	17.25
Unemployed	6241	40.71
Missing	82	

[a] 1 USD ≒ 33 NTD.