Expert Systems with Applications 37 (2010) 598–601

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Applying data mining to explore the risk factors of parenting stress

Hsiang-Yang Chen [a,b], Ting-Wei Hou [a,c], Chao-Hua Chuang [d,*], TBPS Research Group

[a] Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan
[b] Department of Electronic Commerce, Hsing Kuo University of Management, Tainan, Taiwan
[c] Department of Medical Informatics, National Cheng Kung University Hospital, National Cheng Kung University, Tainan, Taiwan
[d] Department of Nursing, Chang Jung Christian University, Tainan County, Taiwan

ARTICLE INFO

Keywords:
Data mining
Decision tree
Parenting stress
Development
Behaviour

ABSTRACT

High parenting stress has been shown to be associated with illness, poor marital relationships, and child abuse. It is thus important to detect and reduce the stress early. Numerous situational, socioeconomic, child, and parent factors contribute to parenting stress. However, regression analysis, the traditional method of exploring risk factors in medicine and social science, has the limitation of not showing the classification, nor exploring unknown potentional factors. Hence, the goal of this study is to explore the risk factors of parenting stress using data mining with decision tree C5.0, to obtain more information. The data are from a professional research group, TBPS, in the National Taiwan University. A total of 206 mother-term born child dyads were recruited to complete the measures of the Parenting Stress Index (PSI), the Child Behaviour Checklist (CBCL), the Comprehensive Developmental Inventory for Infants and Toddlers (CDIIT), and the Chinese Toddler Temperament Scale (CTTS), and so this database includes thousands of variables. The study results indicate that a child development problem, CDIIT, is the major contributing factor to parents with the highest stress, the 90% parenting stress group. For the 80%, 70%, and 60% parenting stress groups, the behavioural problem of children, CBCL, is the major factor causing parenting stress. The data mining decision tree showing the classification route of risk factors is better than the regression model at detecting the significant factors. The findings in this work are considered helpful references for medical staff and social workers to help parents prevent and reduce their parenting stress and thus promote health.

1. Introduction

Parenting stress is concerned with stress in the parent–child system as perceived by the parent. Parenting demands can contribute to the dysfunctional development of the relationships between parent and child, and thus constitute risk factors for the psychopathology of all parties in the system (Abidin, 1995; Deater-Deckard, 1998). The effects of parenting stress are wide-ranging and varied, and such stress has been shown to be associated with illness, poor marital relationships (Belsky, Lang, & Rovine, 1985), and child abuse (Mash, Johnston, & Kovitz, 1983). Hence, quantifying parental stress is an important part of the early detection and intervention to prevent more serious adverse outcomes. Studies show many factors contribute to parental stress, such as the developmental disability of children (Hauser-Cram, Warfield, Shonkoff, & Krauss, 2001), child behavioural problems (Briggs-Gowan, Carter, Skuban, & Horwitz, 2001), infant temperament (Secco & Moffatt, 2003), maternal and child characteristics (Copeland & Harbaugh, 2005; Ong, Chandran, & Boo, 2001), parenting style (Woolfson & Grant, 2006), and children's diseases (Faught, Bierl, Barton, & Kemp, 2007; Hung, Wu, & Yeh, 2004; Macias, Roberts, Saylor, & Fussell, 2006). The relationships between risk factors contributing to parenting stress are thus very complex and ambiguous. Regression analysis, the traditional method of exploring risk factors in medicine and social science, helps detect the significant risk factors among the common suspected factors but it can not show the classification nor explore the unknown factors. Therefore, using a data mining approach to find potentional factors from large amount of data is a promising method (Chang, 2007; Liao, Hsieh, & Huang, 2008), one that may provide us with the indicative predictors of parenting stress, in order to prevent and reduce the adverse effects of such stress as early as possible. As we known, very few studies used data mining methods to explore parenting stress. Hence, the purpose of this study is to explore the risk factors of parenting stress using the data mining, decision tree C5.0, to find more information.

2. Literature review

2.1. Parental Stress Index (PSI)

The Parental Stress Index (PSI) is a commonly used instrument in pediatric settings (Abidin, 1995). Measuring the interrelationship

* Corresponding author. Tel.: +886 6 2785123x3152; fax: +886 6 2785581.
E-mail address: chchuang@mail.cjcu.edu.tw (C.-H. Chuang).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.05.028

between the child's and the parents' characteristics, the PSI has 120 items (54 items related to the parents, 47 to the child, and 19 to life stress). The simplified PSI (PSI/SF) is often used in clinical settings instead of the full version to expedite the administration of the questionnaire. The PSI/SF, consisting of 36 Likert-type items, is adequate in describing the primary components of the parent–child system and has good internal consistency in the Spanish and Chinese versions (Solis & Abidin, 1991; Yeh, Chen, Li, & Chuang, 2001).

2.2. The Child Behaviour Checklist for age 1½–5 (CBCL/1½–5)

The CBCL/1½–5 is a questionnaire based on the parents' report on the children's behavioural, emotional and social functioning at ages 1.5–5 years (Achenbach & Rescorla, 2000). The CBCL/1½–5 consists of 100 items with 99 items for specific behaviour problems and the last one is free for parents to write in children's problems that are not listed in the questionnaire. Each item is scored on a three-point scale: 0 as "not true of the child", 1 as "somewhat or sometimes true" and 2 as "often true or very true." The total problem score is the sum of the scores of 100 items. Seven behavioural syndromes have been identified from 67 of the items in the CBCL/1½–5, namely emotionally reactive (nine items), anxious/depressed (eight items), somatic complaints (11 items), withdrawn (eight items), sleep problems (seven items), attention problems (five items), and aggressive behaviour (19 items).

2.3. The Comprehensive Developmental Inventory for Infants and Toddlers (CDIIT)

CDIIT is a developmental test that is commonly used for infants and toddlers in Taiwan. Five domains of cognition, motor, language, self-help, and social development are constructed. The CDIIT has acceptable test–retest reliability ($r = 0.89 - 0.99$, $p < 0.001$), internal consistency (Cronbach α = 0.75–0.99), content, concurrent and construct validity (Wang, Su, & Liao, 1998; Liao & Pan, 2005; Liao, Wang, & Yao, 2005). Thirty-one items on the CDIIT are scored 0 or 1, indicating failure or success, respectively, during the test or observation at home by the caregivers. From the results of the CDIIT test, the developmental quotient (DQ) of the whole test of the CDIIT of each child is obtained. In the norm, the mean DQ is 100, and the standard deviation (SD) of DQ is 15.

2.4. The Chinese Toddler Temperament Scale (CTTS)

The temperament types are frequently used in clinics and easily explained to parents. The Carey's Toddler Temperament Scale has been applied to Chinese culture and normative data have been established for different genders in 1987 as part of the Chinese Toddler Temperament Scale (CTTS). The scale is a 97-item parent response questionnaire and was standardized on 308 toddlers aged from 12 to 36 months in Taiwan (Tsou, Ju, & Chen, 1987). The main caregivers of the children were asked to rate their children's typical behaviour during the previous on month and items are rated on a six-point Likert scale, ranging from never to always. Three major categories of children's temperament, easy, intermediate, and difficult, are defined according to scores of five temperament dimensions (rhythm, approach/withdrawal, adaptability, mood, intensity) (Carey, 1970). From the results of the CTTS, each child in this study was assigned to one of temperament types (i.e. easy, intermediate, and difficult).

2.5. Data mining

Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand, Mannila, & Smyth, 2001). Knowledge Discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data mining (Frawley, Piatetsky-Shapiro, & Matheus, 1991), and can be used to create predictive and descriptive models. A predictive model makes a prediction about values of data using known results found from different data. Using data mining to generate a predictive model requires classification, regression, time series analysis, and prediction (Dunham, 2003). Such classification techniques are widely used in medical diagnosis, loan approval, detecting faults in industry applications, classifying financial market trend and so on.

2.5.1. Decision tree

A decision tree is a predictive modeling technique used in classification, clustering, and prediction. A decision tree is a diagram, used as a visual and analytical decision support tool where the expected values of competing alternatives are calculated. It is essentially a tree where the root and each internal node are labeled with a question, and arcs then descend from each node to its children, representing answers to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration, and a decision tree is thus a mapping from observations of an item to conclusions of its target value. The decision tree models (algorithms) include ID3 (Iterative Dichotomiser 3), C4.5, C5.0, CART (Classification and Regression Tree), CHAID (Chi-squared Automatic Interaction detection) and QUEST (Quick, Unbiased and Efficient Statistical Tree).

ID3 is a heuristic method for providing a decision tree (Quinlan, 1986). The basic idea is to generate a decision tree by employing a top-down, greedy search through the training data set at each of its tree node, seeking the attribute that best separates the instances. The C4.5 algorithm improves ID3 (Quinlan, 1993) with regard to the splitting rule and the calculation method. C5.0 is a commercial version of C4.5, such as Clementine and RuleQuest (Quinlan, 1997), and it improves the rule generation of C4.5.

CART generates a binary decision tree, unlike ID3 which only creates two children (Breiman, Friedman, Olshen, & Stone, 1984). CART and CHAID provide a set of rules that can be applied to a new (i.e. unclassified) data set to predict which records will have a given outcome. CART segments a data set by creating two-way splits, but CHAID segments a data set by Chi-square tests to create multi-way splits. CART typically requires less data preparation than CHAID. QUEST, another type of decision tree, is similar to the CART algorithm developed by Loh and Shih (1997), but is designed to reduce the processing time required for large CART analyses.

3. Empirical study

3.1. Research structure

Clementine 10.0 is a commercial data mining tool (SPSS, 2005). It supports various decision tree models, such as C5.0, CART (C&RT), Quest, and CHAID. For analyzing the nominal variables, we used the C5.0 tree of Clementine 10.0 to mine the data.

The research procedures are listed below:

1. *Problem definition:* Higher parenting stress can affect the health and relationship between parents and children. Hence, it is important to explore the dominant variables to predict the parenting stress and early intervention to reduce it.
2. *Data collection:* The medical data were collected by the Taiwan Birth Panel Study (TBPS) Research Group when pregnant women went to the National Taiwan University Hospital for