Chung Shan Medical Journal 2022; 33:115-130

Original Article

Access Equity and Life Care Needs among Immigrant Women Using the Target-Association Rules Mining Method

Ming-Hseng Tseng[1†], Chuan-Chao Lin[2,3,†], Chi-Chih Wang[2,4,], Hui-Ching Wu[5,6,*]**

[1] *Department of Medical Informatics, Chung Shan Medical University, Taichung City, Taiwan*
[2] *School of Medicine, Chung Shan Medical University, Taichung City, Taiwan*
[3] *Department of Physical Medicine and Rehabilitation, Chung Shan Medical University Hospital, Taichung City, Taiwan.*
[4] *Department of Physical Medicine and Rehabilitation, Chung Shan Medical University Hospital, Taichung City, Taiwan.*
[5] *Department of Medical Sociology and Social Work, Chung Shan Medical University, Taichung City, Taiwan.*
[6] *Social Service Section, Chung Shan Medical University Hospital, Taichung City, Taiwan.*
[†] *Co-first Author*
[**] *Co-corresponding Author*

Many real world applications of association rule mining from large databases help users make good decisions. However, previous studies produced large numbers of irrelevant patterns, and much time was wasted in finding meaningful rules in large and sparse data sets. This study aims to efficiently discover interesting rules that connote causality between the antecedent and the consequence in a target pattern. In this paper, we propose an improved target-association rule mining method that can remove imprecise patterns, rapidly discover target rules, and apply the method to find associations between socio-economic characteristics and life care needs for new immigrant women in Taiwan. Experimental results show that the proposed method outperforms four other algorithms, namely, Apriori, Apriori-CAR, FP-growth, and DCIP, especially for lower supports. The results also confirm that our approach is practical and effective, with good performance for mining target-association rules in sparsely distributed databases.

Keywords: Equity, Immigrant women, Target-association rule mining, Life care needs

1. Purpose

The Ministry of Internal Affairs of the Republic of China, Taiwan reports that the married immigrant population totaled 41 million from 1987 to 2012. In 2012, the total number of married couples was 142,846, the number of transnational married couples was 20,115 (14.08%), the number of spouses from mainland China, Hong Kong, and Macao was 12,308 (8.62%), the number of spouses from Southeast Asia was 4,718 (3.3%), and the number of spouses from other foreign countries was 3,089 (2.16%). The number of foreign spouses of women was 15,827 (11%), which was higher than the number of foreign spouses of men (4,288, 3%) [1]. Most of the new immigrant women married poor men. The factors of economic distress are often closely associated with immigrant women's health-related quality of life [2]. Immigrants confront problems, such as language, culture, and interpersonal communication. Providing appropriate social support to meet immigrant women's life care needs in Taiwan

* Corresponding Author: Hui-Ching Wu
Address: No. 110, Sec. 1, Jianguo N. Rd., Taichung City, 40201, Taiwan.
Tel: +886-4-24730022 ext. 12137
E-mail: graciewu@csmu.edu.tw

is therefore important.

The algorithms and techniques of data mining (DM), or generally, knowledge discovery in databases (KDD), attempt to find useful patterns that characterize large data sets. However, the development of DM faces the obstacle of large requirements in time and space for processing. These difficulties are inherent in KDD especially because DM techniques are generally applicable and of interest when the target data set is relatively large and sparse [3,4].

Association rule mining is one of the most popular data mining techniques. The task of association rule mining is to extract a set of highly correlated items shared among a large amount of records from a customer database. For example, the rules found from a sales database are useful in the decision making of a marketing manager. The application of these association rules is in market basket analysis. This mining helps the decision maker analyze customers' purchasing habits by determining the associations among items. Agrwal et al. [5,6] first developed the Apriori algorithm to solve the association rule mining problem. At present, association rule mining has been an important direction in database mining. Efficiency and scalability concerns continue to remain significant problems for Apriori-like algorithms [7].

Finding frequent itemsets is the most essential and fundamental work in mining association rules [6,7]. The efficient discovery of frequent itemsets presents a challenge because the number of different combinations of the items in a large database exponentially grows. Furthermore, itemsets may tend to be sparsely distributed, and this makes the counting process difficult. In this study, the immigrant women survey data set has wide and sparse characteristics that make the use of traditional association rule mining algorithms inefficient. To efficiently explore the associations between immigrant women's socio-economic characteristics and their life care needs, an improved target-association rule mining approach is proposed and evaluated in the study.

Related Work

Apriori algorithm [5,6] is the most famous and basic among many of the association rule mining algorithms. The core idea of Apriori algorithm is based on a recursive method of frequent set theory, and its purpose is to identify the association rules whose support and confidence are not less than the given minimum support threshold and minimum confidence threshold, respectively.

Apriori algorithm has two main functions: discover frequent itemsets and generate association rules. It adopts an iterative method to discover frequent itemsets with two steps: the join step and the prune step. In the join step, a candidate *k*-item sets is generated by the joining of two frequent (*k-1*)-itemsets. Then, in the prune step, all itemsets whose (*k-1*)-subset is not a frequent *k*-itemsets are removed from the candidate *k*-itemsets. Finally, the database is scanned to compute the support of the candidate *k*-itemsets. This process is repeated until no new candidate *k*-itemsets is generated[7].

However, two bottlenecks are involved in Apriori algorithm. One is the complex process of candidate generation, which uses much time, space, and memory. Another bottleneck is the multiple scanning of the database. A huge calculation and a complicated transaction process are required by the algorithm. Therefore, the mining efficiency of Apriori algorithm is poor if the transaction database is large, especially for data sets whose distribution is sparse and wide. Many researchers have proposed a number of ways to improve or expand the Apriori algorithm in recent years [7,8].

Han et al. [9] proposed the FP-growth method that uses a prefix-tree data structure to store the frequency information of the original database in a compressed form. Only two database scans are needed for the algorithm, and no candidate generation is required. The FP-growth method has good adaptability to rules of different lengths and has better efficiency than Apriori algorithm. However, it needs to scan the entire database twice, and its efficiency in dealing with large and sparse databases is low.

In recent decades, many investigators have devoted their efforts to high-utility itemset (HUI) mining. Based either on the confidence or utility values, the discovering of high-utility associations rules (HARs) from HUI enables people to select important rules. Mai et al.[10] suggested a method, LNR-HAR algorithm generating all NR-HARs

based on a lattice of HUIs, for the efficient mining of non-redundant high-utility association rules (i.e., NR-HARs).

In general, a large number of patterns will be generated using the traditional association rule mining techniques. To reduce the number of patterns mined and time cost for finding meaningful patterns, Lee et al.[11] proposed the so-called 'target pattern' which has specific associations between antecedent and consequence in a rule. Liu et al. [12] worked on a new method for mining class-association rules, and it is called classification based on associations. The target attribute (or class attribute) is not pre-determined for association rule mining. However, the target attribute must be pre-determined in classification problems. Some algorithms for mining classification rules on the basis of association rule mining have been proposed [13].

Recently, Huang presented DCIP algorithm [14]. In this algorithm, a data cutting and vector inner product method is used to quickly generate candidate itemsets and count support, respectively. Huang reported that DCIP is faster and more efficient than Apriori while achieving the same level of accuracy. Although it uses an efficient vector inner operator, DCIP can still suffer from the high cost of target-association candidate generation. To address this problem, a new algorithm based on a useful Boolean matrix [15,16] and bit-wise operations [17,18] are proposed for the efficient mining of target-association rules in this paper.

2. Materials and Methods

2.1. Data Source

This study employed secondary survey data reported in "The Survey of Living Conditions of Foreign and Mainland China Spouses 2008" (National Immigration Agency 2008) from the Ministry of the Interior, Republic Of China. The respondents of the survey project provided their basic personal information, such as status of family members, employment, health care needs, life adaptation, and overall living environment experience, for a total of 48 variables. The number of the target population was 407,487. The sampling method was a proportional systematic sampling method. The number of completed survey samples was 13,345 respondents.

The study protocol was approved by the Institutional Review Board of Chung Shan Medical University Hospital (permission no. CS13049). Finally, 8,424 records with 15 attributes were employed in the following analysis for immigrant women's life care needs; data processing techniques were used to exclude missing data and outliers.

2.2. Basic Concept

The association rule mining problem can be formally stated as follows: Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of m-dimensional attribute values, called items. Let $D = \{T_1, T_2, \ldots, T_n\}$ be a set of tuples with cardinality n, called the transaction database, where each transaction (tuple) $Ti \in D$ has a set of items, called an itemset, such that $Ti \subseteq I$. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k-itemsets. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. X is called the antecedent whereas Y is called the consequent.

An association rule $X \rightarrow Y$ can have different measures that denote its significance and quality. To measure whether the identified rules are meaningful, three quality measurements are employed in this study, namely, support, confidence (reliability), and lift (correlation).

The general idea is that if, say, XY is frequently 2-itemsets, then we can determine if the association rule $X \rightarrow Y$ is meaningful by calculating support, confidence, and lift. Support ($X \rightarrow Y$) is the percentage of the numbers in the transaction set D containing $X \cup Y$, and it implies the frequency of occurring patterns. Confidence ($X \rightarrow Y$) is defined as the fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X, and it means the strength of the association rules. Lift ($X \rightarrow Y$) measures the correlation between 2-itemsets XY, and it characterizes the direction of the relationship between the antecedent X and the consequent Y of the association rule.

The support of the rule $X \rightarrow Y$ is defined in (1), and its confidence and lift are defined in (2) and (3),

respectively:

$$Support(X \rightarrow Y) = P(X \cup Y) \quad (1)$$

$$Confidence(X \rightarrow Y) = P(Y|X) = \frac{P(X \cup Y)}{P(X)} = \frac{Support(X \rightarrow Y)}{Support(X)} \quad (2)$$

$$Lift(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{Support(X \rightarrow Y)}{Support(X) \cdot Support(Y)} = \frac{Confidence(X \rightarrow Y)}{Support(Y)} \quad (3)$$

where $P(X \cup Y)$ is the percentage of transactions in D that contain $X \cup Y$ and support(X) is defined as the percentage of transactions that containing X.

A set of items X is said to be frequent if and only if support (X) is greater than the user-defined minimum support. Therefore, X is a frequent 1-itemset. If the rule $X \rightarrow Y$ has support greater than the minimum support, its confidence is greater than the minimum confidence, and its lift is bigger than 1, then $X \rightarrow Y$ is an interesting rule, that is, an association rule. From (1), (2), and (3), finding the association rules is effectively equivalent to generating all the frequent itemsets with support greater than the minimum support. Therefore, we focus on frequent itemset mining.

Rules that have a support value greater than the user-defined minimum support (*min_support*), in which the itemset needs to be present in the minimum threshold number of transactions, and confidence greater than the user-defined minimum confidence (*min_confidence*) are called valid association rules. The lift symbolizes whether the association is positive or negative. A lift value greater than 1 indicates a positive relationship between the itemsets, a lift value less than 1 indicates a negative relationship, and a lift value equal to 1 indicates that the itemsets are independent and that no relationship exists between the itemsets.

2.3. New Algorithm Description

Within this framework, we consider mining Boolean target-association rules. The proposed algorithm can be divided into two steps. Initially, the algorithm finds all frequent itemsets with the target item (consequent). Then, it generates all target-association rules from frequent itemsets. Obviously, the second step is easier than the first step, so we will only describe the algorithm to discover the frequent itemsets.

Fig.1 provides a pseudo-code description of

Input: transaction database D, minimum threshold of support *min_support*
Output: frequent itemsets L_Y

(1) scan D once, generate Boolean matrix R, and delete infrequent items;
(2) for all candidate $c \in I \subset R$ do begin
(3) compute $c.count = Support(AV_j)$ by using Eq. (4)
(4) $L_1 = L_1 \cup c \mid c.count \geq min\_support$
(5) end for
(6) selected consequent item Y, delete TV_i if Y=0 and $L_{1Y} = L_1$;
(7) for (k=2; $L_{(k-1)Y} \neq \emptyset$; k++) do begin
(8) count $Sum(TV_i)$ by using Eq. (5) and delete TV_i if $Sum(TV_i) < k$;
(9) generate C_{kY} by combining any (k-1) items in $L_{(k-1)Y}$ with the consequent item Y;
(10) for all candidate $c \in C_{kY}$ do begin
(11) compute $c.count = Support(\vec{X}_1, \vec{X}_2, \cdots \vec{X}_{k-1}, \vec{Y})$ by using Eq. (8)
(12) $L_{kY} = L_{kY} \cup c \mid c.count \geq min\_support$
(13) end for
(14) end for
(15) Answer $L_Y = \cup_k L_{kY}$

Fig. 1. A pseudo-code of the proposed algorithm.

airiti

Table 1. Example database

TID	Items
T1	A, B, E
T2	E
T3	B, C
T4	A, D, E
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

the proposed algorithm. A detailed example of the proposed approach is given as follows.

Consider an example database shown in Table 1. Five different items (attributes) and nine transactions (tuples) are presented. Suppose that the minimum support number for this example is two (*min_support* = 2).

Step 1: Generate the Boolean matrix R.

At first, the transaction database D can be mapped into a Boolean matrix [15,16] $R = (r_{ij})_{nxm}$ with n rows (transactions) and m columns (items). The absence or presence of an item is represented as 0 or 1 for each transaction $T_i \in D$. Transactions are strings of 0s and 1s. To determine if a particular itemset is frequent, we count the number of records where the values for all the items in the itemset are 1. A corresponding Boolean matrix R will be determined after the scanning of database D once. Therefore, the issues of association rule mining can be translated into the analysis of Boolean matrix. Each row in R corresponds to a relevant transaction T_i called tuple bit-vector TV_i, $1 \le i \le n$. Each column in R corresponds to a relevant item I_j, called attribute bit-vector AV_j, $1 \le j \le m$. The Boolean matrix R can be expressed as $R = (r_{ij})_{nxm} = (TV)_{i=1\sim n} = (AV)_{j=1\sim m}$.

Table 2 lists the candidate1-itemsets C_1 based on the mapped Boolean matrix R. The last row in Table

Table 2. Boolean matrix C_1 of example database

	A	B	C	D	E
T1	1	1	0	0	1
T2	0	0	0	0	1
T3	0	1	1	0	0
T4	1	0	0	1	1
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0
S_{upport}	6	5	6	1	4

Table 3. Boolean matrix L_1 of example database

	A	B	C	E	S_{um}
T1	1	1	0	1	3
T2	0	0	0	1	1
T3	0	1	1	0	2
T4	1	0	0	1	3
T5	1	0	1	0	2
T6	0	1	1	0	2
T7	1	0	1	0	2
T8	1	1	1	1	4
T9	1	1	1	0	3

2 shows the support number of each item calculated by Eq. (4). We generate the resulting $C_1 = \{A, B, C, D, E\}$.

Step 2: Generate the frequent 1-itemsets L_1.

We count the support number of every attribute bit-vector AV_j, which is expressed as $Support(AV_j)$. At the same time, we also count the sum of every tuple bit-vector TV_i, which is expressed as $Sum(TV_i)$. If the support number of AV_j is lower than the user specified threshold *min_support* (*i.e.,* $Support(AV_j) <$ *min_support*), then we delete all infrequent items to generate the frequent 1-itemsets L_1.

$$Support(AV_j) = \sum_{i=1}^{n} r_{ij} \quad (4)$$

$$Sum(TV_i) = \sum_{j=1}^{m} r_{ij} \quad (5)$$

From Table 2, we delete those items lower than the minimum support to avoid redundant data scanning. In this example, item D is deleted from Table 2 because its support number is lower than 2. This step generates the frequent 1-itemsets, as shown in Table 3. Therefore, the resulting L_1 is $\{A, B, C, E\}$. The last column on the right side of Table 3 shows the sum of all items for each transaction in L_1 with the use of Eq. (5).

Step 3: Generate the candidate 2-itemsets C_{2Y} for consequent item Y.

After selecting the consequent item (target class Y), we remove all $Y=0$ transactions and some tuple bit-vectors TV_i where $Sum(TV_i)$ is lower than 2 to find the candidate 2-itemstes C_{2Y} for the consequent item Y. This step deletes unnecessary transactions to reduce the data size to be analyzed once more.

After deciding the target class $Y = E$ in Table 3, we delete all $E=0$ transactions, such as T3, T5, T6, T7, and T9. At the same time, we also delete transaction T2 because of its $Sum(TV_2)$, which is lower than 2, to find the candidate 2-itemstes C_{2E}. The elements in C_{2E} can be obtained by combination of each item of L_1 with the consequent item E to generate $C_{2E} = \{AE, BE, CE\}$, as shown in Table 4.

Step 4: Generate the frequent 2-itemsets L_{2Y} for consequent item Y.

Table 4. Boolean matrix C_{2E} of example database

	$A:(X_1)$	$B:(X_2)$	$C:(X_3)$	$E:(Y)$
T1	1	1	0	1
T4	1	0	0	1
T8	1	1	1	1

Table 5. Boolean matrix L_{2E} of example database

	$A:(X_1)$	$B:(X_2)$	$E:(Y)$	S_{um}
T1	1	1	1	3
T4	1	0	1	2
T8	1	1	1	4

Let $\vec{X}$ represent an attribute bit-vector AV_j in C_{2Y}, $\vec{Y}$ is the target bit-vector in C_{2Y}, and $\vec{X} \neq \vec{Y}$. The support of two bit-vectors $\vec{X}$ and $\vec{Y}$ with cardinality n_2 is defined as

$$Support(\vec{X}, \vec{Y}) = \sum_{i=1}^{n_2} AND(x_i, y_i) \quad (6)$$

where the *AND* operator is a bitwise operation whose result is 1 if x_i is 1 and y_i is 1; otherwise, the result is 0. Determining if the 2-itemset *XY* is frequent is thus reduced to testing if $Support(\vec{X}, \vec{Y}) \geq min\_support$. If the support number of *XY* is lower than the user-specified threshold *min_support*, then we delete the 2-itemset *XY* to generate the frequent 2-itemsets L_{2Y}. We also count the sum of every remaining tuple bit-vector TV_i in L_{2Y}.

From Table 4, four attribute bit-vectors AV_j are found in C_{2E}, and they are $\vec{A}$ = (1, 1, 1), $\vec{B}$ = (1, 0, 1), $\vec{C}$ = (0, 0, 1), and $\vec{E}$ = (1, 1, 1). We can count the support number by using Eq. (6) for $Support(\vec{A}, \vec{E}) = 3$, $Support(\vec{B}, \vec{E}) = 2$, and $Support(\vec{C}, \vec{E}) = 1$. For example, $Support(\vec{B}, \vec{E}) = \sum_{i=1}^{3} \text{AND}(b_i, e_i) = 1 + 0 + 1 = 2$. Because $Support(\vec{C}, \vec{E}) = 1 < min\_Support$, we delete the 2-itemset *CE* to generate the frequent 2-itemsets $L_{2E} = \{AE, BE\}$, as shown in Table 5. The last column on the right side of Table 5 represents the sum of all items for each transaction in L_{2E} with the use of Eq. (5).

Step 5: Generate the candidate 3-itemsets C_{3Y} for consequent item Y.

We delete some tuple bit-vectors TV_i where $sum(TV_i)$ is lower than 3 to find the candidate 3-itemstes C_{3Y} . The elements in C_{3Y} are obtained by combination of any two items with the consequent item *Y* in L_{2Y}.

From Table 5, we delete transaction T4 because its $sum(TV_4)$ is lower than 3 to find the candidate 3-itemstes C_{3E}. The elements in C_{3E} can be obtained by combination of any two items of L_{2E} with the consequent item *E* to generate $C_{3E} = \{ABE\}$, as shown in Table 6.

Step 6: enerate the frequent 3-itemsets L_{3Y} for consequent item Y.

Let $\vec{X}_1$ and $\vec{X}_2$ represent an attribute bit-vector AV_j in C_{3Y}; $\vec{Y}$ is the target bit-vector in C_{3Y}, and $\vec{X}_1 \neq \vec{X}_2 \neq \vec{Y}$. The support of three bit-vectors $\vec{X}_1$, $\vec{X}_2$, and $\vec{Y}$ with cardinality n_3 is defined as

$$Support(\vec{X}_1, \vec{X}_2, \vec{Y}) = \sum_{i=1}^{n_3} AND(x_{1i}, x_{2i}, y_i) \quad (7)$$

where the *AND* operator is a bitwise operation whose result is 1 if $x_{1i} = x_{2i} = y_i = 1$; otherwise, the result is 0. Determining if the 3-itemset X_1X_2Y is frequent is thus reduced to testing if $Support(\vec{X}_1, \vec{X}_2, \vec{Y}) \geq min\_support$. If the support number of X_1X_2Y is lower than the user-specified threshold *min_support*, then we delete the 3-itemset X_1X_2Y to generate the frequent 3-itemsets L_{3Y}. We also count the sum of every remaining tuple bit-vector TV_i in L_{3Y}.

The above process is repeated with successively increasing number *k* until either C_{kY}, or L_{kY} is empty. At the end of procedure, we can determine the all-frequent itemsets. The support of *k* bit-vectors $\vec{X}_1$, $\vec{X}_2$,$\cdots\vec{X}_{k-1}$ and $\vec{Y}$ with cardinality n_k is defined as

$$Support(\vec{X}_1, \vec{X}_2, \cdots \vec{X}_{k-1}, \vec{Y}) = \sum_{i=1}^{n_k} AND(x_{1i}, x_{2i}, \cdots x_{(k-1)i}, y_i) \quad (8)$$

From Table 6, three attribute bit-vectors AV_j are found in C_{3E}, and they are $\vec{A}$ = (1, 1), $\vec{B}$ = (1, 1), and $\vec{E}$

Table 6. Boolean matrix C_{3E} of example database

	$A:(X_1)$	$B:(X_2)$	$E:(Y)$
T1	1	1	1
T8	1	1	1

= (1, 1). We can count the support number by using Eq. (7) for $Support(\vec{A}, \vec{B}, \vec{E}) = 2$. Because $Support(\vec{A}, \vec{B}, \vec{E}) \geq min\_support$, we can find the frequent 3-itemsets $L_{3E} = \{ABE\}$. Finally, we stop repeating the above process because C_{4E} is empty. At the end of the procedure, we can determine the all-frequent itemsets $L_{2E} = \{AE, BE\}$ and $L_{3E} = \{ABE\}$ from the example database.

With the use of Eqs. (1)~(3), the values of *Support*($A \rightarrow E$), *Confidence*($A \rightarrow E$), and *Lift*($A \rightarrow E$) are 3, 0.5, and 4.5, respectively. Furthermore, the values of *Support*($AB \rightarrow E$), *Confidence*($AB \rightarrow E$), and *Lift*($AB \rightarrow E$) are 2, 0.67, and 1.5, respectively. These results show that both *Rule*($A \rightarrow E$) and *Rule*($AB \rightarrow E$) are meaningful. By contrast, *Rule*($B \rightarrow E$) is not interesting because *Lift*($B \rightarrow E$)=0.9, although *Support*($B \rightarrow E$)=2 and *Confidence*($B \rightarrow E$)=0.4.

3. Results

3.1. Descriptive Statistics

Table 7 shows the life care needs for all the participants (13,345). In the theme "life care needs", the questionnaire items included "protect employment equity", "provide living assistance measures", "assist children in school", "establish dedicated service agencies", "set up an advisory services window", "provide childcare for the children", "increase multicultural activities", and "increase life adaptation assistance". The top three items of life care needs for new immigrant women were "protect employment equity" (22.28%), "provide living assistance measures" (11.22%), and "assist children in school" (9.29%). A total of 8,424 records were used in the following analysis, excluding data designated as "other" or "missing value."

All categorical variables were compared with Chi-Square χ^2 test for the different life care needs. A p-value < 0.05 indicates statistical significance. Table 8 presents 15 significant characteristics; for example, 3,072 (36.47%) came from "Southeast Asia, Hong Kong, Macao, and other countries," and 5,352 (63.53%) came from the "Mainland China." When asked about the "type of identity documents" they have, 3,001 (35.62%) said they were "foreign spouses", and 5,423 (64.38%) said they were "Mainland China spouses". When asked about their "main source of income", 5,164 (61.30%) said it was "provided by family members (husband, parents of the husband, children)", and 3,001 (35.63%) said it was "from [their] work income and savings". When they were asked the question, "does your husband have a job?", 6,480 (76.92%) answered "yes" and

Table 7. Statistical percentage of 'life care needs' among Taiwan new immigrant women

Items of life care needs		Number	%
Protect employment equity		2,973	22.28
Provide living assistance measures		1,497	11.22
Assist children in school		1,240	9.29
Establish dedicated service agencies		756	5.67
Set up an advisory services window		681	5.10
Provide childcare for the children		526	3.94
Increase multicultural activities		429	3.21
Increase life adaptation assistance		322	2.41
	Subtotal	8,424	63.12
Missing value (contains other)		4,921	36.88
	Total	13,345	100.00

Data Source: The Survey of Living Conditions of Foreign and Mainland China Spouses 2008, the Ministry of the Interior, R.O.C.

Table 8. Descriptive statistics of new immigrant women

Characteristics	*N=8,424*	*%*	*P(χ^2)*
Nationality			0.000*
Southeast Asia, Hong Kong, Macao, Other countries	3072	36.47	
Mainland China	5352	63.53	
Education degree			0.000*
Illiterate, self-learning, primary school	1813	21.52	
Junior high school	3220	38.22	
Senior high school	2465	29.26	
College, University, Graduate	895	10.63	
Missing value	31	0.37	
Type of identity documents			0.000*
Foreign spouses	3001	35.62	
Mainland China spouses	5423	64.38	
Duration of residence in Taiwan			0.000*
Less than 4 years	1694	20.11	
4 years to less than 6 years	1460	17.33	
6 years to less than 8 years	1858	22.06	
More than 8 years	3407	40.44	
Missing value	5	0.06	
Marriage frequency			0.000*
The first time	7322	86.92	
More than twice	1055	12.52	
Missing value	47	0.56	
The main source of pocket money			0.000*
No	162	1.92	
Working income and savings	3001	35.63	
Provided by husband, parents of husband, children	5164	61.30	
Missing value	97	1.15	
Education degree of husband			0.000*
Illiterate, self-learning, primary school	1131	13.43	
Junior high school	2341	27.79	
Senior high school	3443	40.87	
College, University, Graduate	1418	16.83	
Missing value	91	1.08	

Characteristics	*N=8,424*	*%*	*P(χ^2)*
Marriage frequency of husband			0.000*
The first time	6722	79.8	
More than twice	1670	19.82	
Missing value	32	0.38	
Does husband have a job?			0.000*
Yes	6480	76.92	
No	1936	22.98	
Missing value	8	0.1	
Average income of husband			0.000*
0~ NT$ 19,999 (US$ 666)	954	11.33	
NT$ 20,000~NT$ 29,999 (US$ 667~US$999)	2218	26.32	
NT$ 30,000~NT$ 39,999(US$ 1000~US1332)	1830	21.72	
Over NT$ 40,000 (US$ 1333)	1286	15.27	
Missing value	2136	25.36	
The number of births with a Taiwanese spouse			0.000*
None	2036	24.17	
One child	2843	33.75	
Two children	3014	35.78	
More children	531	6.30	
Missing value	0	0	
The main source of funds for family living expenses			0.000*
By myself or my husband	8023	95.24	
Provided by relatives, children, NGO, government allowance	312	3.70	
Missing value	89	1.06	
People around me are very friendly			0.000*
Disagree	513	6.09	
Agree	5962	70.77	
Very agree	1946	23.10	
Missing value	3	0.04	
I live in a safe place, disasters and accidents are rare			0.000*
Disagree	566	6.72	
Agree	6131	72.78	
Very agree	1725	20.48	
Missing value	2	0.02	

Characteristics	N=8,424	%	$P(\chi^2)$
Overall, I am very happy in Taiwan			0.000*
Disagree	1198	14.22	
Agree	5685	67.49	
Very agree	1538	18.26	
Missing value	3	0.03	

*$p<0.001$

1,936 (22.98%) said "no". When asked about "number of births with a Taiwanese spouse", 2,036 (24.17%) answered "none", 2,843 (33.75%) answered "one child", 3,014 (35.78%) answered "two children", and 531 (6.30%) answered "more children". When asked about their "main source of funds for family living expenses", 8,023 (95.24%) answered "myself or my husband", and only 312 (3.70%) answered "provided by relatives, children, NGOs, government allowance". Clearly, the statistics shown in Table 2 demonstrate that the life care needs (Y) of the new immigrant women, with more than 95% confidence, significantly relate with all of 15 socio-economic characteristics (X_i) with the use of the Chi-square χ^2 test. Tables 7 and 8 indicate that most of the itemsets were sparsely distributed, so the counting process in this immigrant women survey data set is difficult. The use of traditional association rule mining algorithms will be inefficient.

3.2. Performance Evaluation

We compared the performances of the proposed method with the Apriori, Apriori-CAR, FP-growth, and DCIP algorithms [14]. The proposed algorithm, along with DCIP, was implemented in VC++. The algorithms Apriori, Apriori-CAR, and FP-growth were also compared with Weka software [4]. To reflect the algorithmic runtime performance, all of the figures use computing time only as the performance metric in this study, whereas the time to load and prepare the data is not considered here. All the experiments were conducted on an Intel i7-2637M 1.70GHz machine with 4GB RAM on a WIN7 64 bit platform.

Fig.2 shows the results of the comparison of the proposed method with the Apriori, Apriori-CAR, FP-growth, and DCIP algorithms with different support counts on the data set of the immigrant women's life care needs. The cleaning data set contains 47 items and 8,424 records. A few items co-occurred together in a sparse data set when the support threshold is high. Therefore, FP-growth demonstrates the best performance among the five algorithms for higher supports ($min\_support \geq 0.1$) on this sparse data set, as shown in Fig.2. However, note that as the support drops ($min\_support \leq 0.05$) and the itemsets become long, the proposed method outperforms the other four

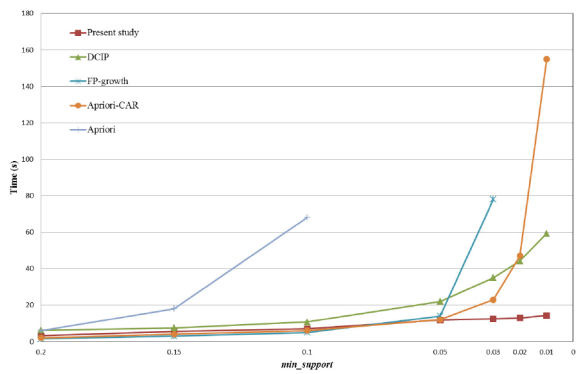

Fig. 2. Execution time required by the proposed method, Apriori, Apriori-CAR, FP-growth and DCIP in different minimum support threshold.

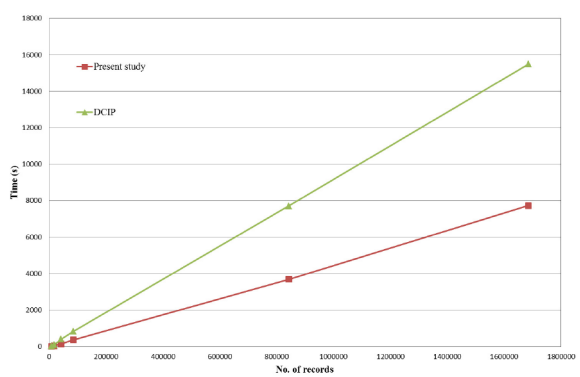

Fig. 3. Execution time required by the proposed method and DCIP in different data sizes.

algorithms because it becomes the fastest algorithm. In addition, Apriori does not work when *min_support* ≤ 0.05, and FP-growth also fails when *min_support* ≤ 0.02. In general, most of the association rule mining algorithms quickly lose their lead once the support level becomes small. The result confirms that both the traditional Apriori and FP-growth algorithms are inefficient in dealing with sparse or large databases. The proposed method clearly performs best when the itemsets are long at low supports. For example, the speedup ratio for the proposed method: Apriori-CAR: DCIP: FP-growth is 5.75: 3.39: 2.16: 1.00 at *min_support* = 0.03. The speedup ratio for the proposed method: DCIP: Apriori-CAR is 10.03: 2.57: 1.00 at *min_support* = 0.01.

To understand the performance of the proposed method in large data sets, Fig.3 shows the results with different data sizes in comparing the proposed method with Apriori-CAR and DCIP from 8,424 records to 1,684,800 (=200*8,424) records, with *min_support* = 0.03. Fig.3 demonstrates the proposed method is more efficient than Apriori-CAR and DCIP in all of the data sizes. Generally, the proposed method is two to three times faster than DCIP.

According to the description in Section of "New Algorithm Description", we employ Boolean matrix forms to efficiently represent the transaction database for mining association rules in the proposed method; both infrequent items and irrelevant transactions are continuously deleted to reduce the large number of recursive scans. Furthermore, the techniques of pre-selecting the consequent item to find specific target rules and using bitwise operators to count supports also largely increase calculation speed. All of the skills described above effectively make the proposed method enhance association rule mining efficiency.

3.3. Life Care Needs Models of Immigrant Women

This study applied the proposed method to perform associations between the socio-economic characteristics and life care needs of new immigrant women in Taiwan. High-frequency items were included in the same target class to facilitate calculation of the association between these characteristics and the specific life care need. This study focused on rules with greater importance than others, so it examined only those with support ≥ 3% and Lift ≥1. The characteristics of extracted target-association rules derived by the proposed method are listed in Table 9 for the top 3 life care needs. From

Table 9. Characteristics of discovering class-association rules between life care needs (*Y*) and socioeconomic status conditions (X_i)

Needs	*k*-itemset	No. of rules	Support		Confidence		Lift	
			min	max	min	max	min	max
Protect employment equity	2	20	0.024	0.338	0.354	0.418	1.003	1.184
	3	270	0.020	0.270	0.353	0.482	1.000	1.365
	4	1241	0.020	0.234	0.353	0.518	1.000	1.468
	5	1722	0.030	0.191	0.353	0.525	1.000	1.487
Provide living assistance measures	2	16	0.035	0.134	0.179	0.325	1.004	1.826
	3	80	0.030	0.117	0.178	0.344	1.002	1.934
	4	144	0.030	0.089	0.178	0.344	1.001	1.934
	5	107	0.030	0.072	0.178	0.335	1.001	1.882
Assist children in school	2	20	0.033	0.142	0.149	0.212	1.009	1.438
	3	136	0.030	0.136	0.147	0.221	1.001	1.498
	4	389	0.030	0.121	0.148	0.227	1.002	1.542
	5	591	0.030	0.105	0.148	0.230	1.003	1.562

Table 10. Top one association rule between life care needs (Y) and socioeconomic status conditions (X_i)

X_1	X_2	X_3	X_4	Y	Support	Confidence	Lift
Nationality = Mainland China	The main source of pocket money = Working income and savings	The number of births with a Taiwanese spouse = None	The main source of family living expenses = By myself or husband	Protect the employment equity	0.032	0.525	1.487
Nationality = Mainland China	Type of identity documents = Mainland China spouses	Does husband have a job? = No	I live in a safe place, disasters and accidents are rare = Agree	Provide living assistance measures	0.045	0.334	1.882
The main source of pocket money = Provided by husband, parents of husband, children	The number of births with Taiwan spouse = Two children	People around me are very friendly = Agree	I live in a safe place, disasters and accidents are rare = Agree	Assist children in school	0.032	0.230	1.562

the results in Table 9, the total numbers of interesting rules for life care need = "protect employment equity," "provide living assistance measures," and "assist children in school" is 3,253, 347, and 1,136, respectively.

Table 10 shows the top association rule ($X_i \rightarrow Y$) with the highest confidence for the three life care needs (Y) with socioeconomic status conditions (X_i). The corresponding confidence of the target rule for life care need = "protect employment equity," "provide living assistance measures," and "assist children in school" is 0.5248, 0.3345, and 0.23, respectively. The value of confidence is not high because the data were sparsely distributed in this study. Only the most important rules with confidence $\geq$ 50% were considered; after the redundant rules were cleaned, 11 rules were found, as shown in Table 11. The results in Table 11 indicate that the detail associations ($X_i \rightarrow Y$) between socioeconomic status conditions (X_i) and life care need (Y) = "protect employment equity".

By analyzing the target-association rules in Table 11, we can see that the subgroup of immigrant women in Taiwan with the most important characteristic is "number of births with Taiwanese spouse = none", that with the second most important characteristic is "nationality = Mainland China" or "type of identity documents = Mainland China spouses", that with the third most important characteristic is "main source of income = work income and savings", and that with the fourth most important characteristic is "main source of funds for family living expenses = myself or my husband".

4. Conclusion

In this paper, we have proposed a new method for mining target-association rules in large and sparse transaction databases. This method has several advantages. First, it uses Boolean matrix representation to compress the database with one scan. Second, it continues to delete both infrequent items and irrelevant transactions during the computation procedures to reduce the large number of recursive scans. Third, it uses a pre-selected target item in the consequent part to remove many imprecise patterns and quickly discover target rules. Finally, it employs bit-vector operators for the efficient counting of support in the entire mining process. We compare our method with the algorithms Apriori, Apriori-

Table 11. Eleven interesting association rules between socioeconomic status conditions (X_i) and life care need (Y) = 'protect the employment equity'

X_1	X_2	X_3	X_4	Y	Support	Confidence	Lift
Nationality = Mainland China	The main source of pocket money = Working income and savings	The number of births with a Taiwanese spouse = None	The main source of family living expenses = By myself or my husband	Protect the employment equity	0.0315	0.5248	1.4869
Type of identity documents = Mainland China spouses	The main source of pocket money = Working income and savings	The number of births with a Taiwanese spouse = None	The main source of family living expenses = By myself or my husband	Protect the employment equity	0.0323	0.5231	1.4821
Nationality = Mainland China	The main source of pocket money = Working income and savings	The number of births with a Taiwanese spouse = None	Type of identity documents = Mainland China spouses	Protect the employment equity	0.0324	0.5180	1.4678
Nationality = Mainland China	The main source of pocket money = Working income and savings	The number of births with a Taiwanese spouse = None	Null	Protect the employment equity	0.0324	0.5180	1.4678
Type of identity documents = Mainland China spouses	The main source of pocket money = Working income and savings	The number of births with a Taiwanese spouse = None	Null	Protect the employment equity	0.0332	0.5166	1.4638
Nationality = Mainland China	Does husband have a job? = Yes	The number of births with a Taiwanese spouse = None	Type of identity documents = Mainland China spouses	Protect the employment equity	0.0402	0.5144	1.4576
Nationality = Mainland China	Does husband have a job? = Yes	The number of births with a Taiwanese spouse = None	Null	Protect the employment equity	0.0402	0.5144	1.4576
Nationality = Mainland China	Does husband have a job? = Yes	The number of births with a Taiwanese spouse = None	The main source of family living expenses = By myself or my husband	Protect the employment equity	0.0391	0.5109	1.4475

X_1	X_2	X_3	X_4	Y	Support	Confidence	Lift
Type of identity documents = Mainland China spouses	Does husband have a job? = Yes	The number of births with a Taiwanese spouse = None	Null	Protect the employment equity	0.0411	0.5081	1.4396
The main source of pocket money = Working income and savings	Does husband have a job? = Yes	The number of births with a Taiwanese spouse = None	Null	Protect the employment equity	0.0240	0.5063	1.4345
Type of identity documents = Mainland China spouses	Does husband have a job? = Yes	The number of births with a Taiwanese spouse = None	The main source of family living expenses = By myself or my husband	Protect the employment equity	0.0399	0.5045	1.4295

CAR, FP-growth, and DCIP, and we determine that the proposed algorithm has better performance than the other four algorithms, especially in situations with low support for sparse data sets.

This study also applied the proposed method to discover the associations between socio-economic characteristics and life care needs for new immigrant women in Taiwan. This research first focused on important rules with support ≥ 3% and Lift ≥1, and it identified that the total number of interesting rules for life care need = "protect employment equity", "provide living assistance measures", and "assist children in school" is 3,253, 347, and 1,136, respectively. Only the most important rules with confidence ≥ 50% are considered in this paper. Finally, a total of 11 rules are found. By analyzing these target-association rules, we can determine that the subgroup of immigrant women in Taiwan who has the primary demand for life care need are "protect employment equity", and they have some important characteristics, such as "number of births with a Taiwanese spouse = none", "nationality = Mainland China" or "type of identity documents = Mainland China spouses", "main source of income = work income and savings", and "main source of family living expenses = myself or my husband".

The results can guide the Taiwanese government in introducing a practical and multicultural immigration policy that will not only help foreign spouses adapt well to society but also cultivate a public sense of respect for the co-existence of different cultures on the path toward the realization of a multi-cultural society. Although limited in scope, this study can serve as a model for future studies that use large data sizes in determining the hidden links between socio-economic characteristics and life care needs for new immigrant women in the world.

Acknowledgments

This work was partially supported by the Ministry of Science and Technology in Taiwan, under grant number MOST 109-2410-H-040-003-SS2. This support is greatly appreciated. The Survey Research Data Archive, Academia Sinica is responsible for the data distribution. The authors appreciate the assistance in providing data by the institutes and Dr. Thung-Hong Lin aforementioned. We thank Mr. Fang-Pin Lioa for performing the data processing.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Reference

1. Department of Household Resistration. 2013. (Accessed 1 June 2020, at http://www.ris.gov.tw/en/web/ris3-english/home.)
2. Singh GK, Rodriguez-Lainz A, Kogan MD. Immigrant Health Inequalities in the United States: Use of Eight Major National Data Systems. The Scientific World Journal 2013;2013.
3. Han J, Kamber M. Data mining: concepts and techniques: Morgan Kaufmann; 2006.
4. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques: Elsevier; 2011.
5. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record; 1993: ACM. p. 207-16.
6. Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc 20th Int Conf Very Large Data Bases VLDB 1994: Citeseer. p. 487-99.
7. Kotsiantis S, Kanellopoulos D. Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering 2006;32:71-82.
8. Maragatham G, Lakshmi M. A Recent Review on Association Rule Mining. Computer Science & Engg Department, Sathyabama University, Chennai, Tamil Nadu, India 2012.
9. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM SIGMOD Record; 2000: ACM. p. 1-12.
10. Mai T, Nguyen LT, Vo B, Yun U, Hong T-P. Efficient algorithm for mining non-redundant high-utility association rules. Sensors 2020;20:1078.
11. Lee DG, Ryu KS, Bashir M, Bae J-W, Ryu KH. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. Journal of medical systems 2013;37:1-10.
12. Liu Bing HW, Ma Yiming. Integrating classification and association rule mining. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), AAAI Press, 1998 1998; New York City, NY, United States. p. pp. 80-6.
13. Nguyen LT, Vo B, Hong T-P, Thanh HC. CAR-Miner: An efficient algorithm for mining class-association rules. Expert Systems with Applications 2013;40:2305-11.
14. Huang YC. The application of data mining to explore association rules between metabolic syndrome and lifestyles. Health Information Management Journal 2013;42:29.
15. Wur S-Y, Leu Y. An effective Boolean algorithm for mining association rules in large databases. Database Systems for Advanced Applications, 1999 Proceedings, 6th International Conference on; 1999: IEEE. p. 179-86.
16. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 2nd ed: Morgan Kaufmann; 2011.
17. Dunkel B, Soparkar N. Data organization and access for efficient data mining. Data Engineering, 1999 Proceedings, 15th International Conference on; 1999: IEEE. p. 522-9.
18. Song W, Yang B, Xu Z. Index-BitTableFI: An improved algorithm for mining frequent itemsets. Knowledge-Based Systems 2008;21:507-13