H-1B AND PERMANENT VISA APPLICATION CASES ANALYSIS

Data Mining Report (Part VI)

ABSTRACT

Using Apriori algorithm to create a set of association rules for association analysis.

HUA, MENGLI/ MA, LIANG

CSE7331 Data Mining

Contents

II. Executive Summary

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.[1] There are two kinds of relationships in relational analysis: simple associations and sequence associations. The famous implement of simple relationship is the shopping basket analysis. An example of a classic shopping basket analysis says that 80% of customers buying bread buy milk. The combination of bread and milk as a kind of breakfast is accepted by everyone. There is no common attribute between the two, but the combination is a delicious breakfast, which is a simple relationship.

Another example for sequence relationship, 80% of customers who buy iphone mobile phones will choose to buy iphone mobile phone protection shell, this is the sequence of relationships, generally no one to buy a protective shell and then buy a mobile phone. This is a sequential order of time.

Simple association rules are unguided learning methods that focus on exploring internal structures.[2] Simple association rules are also the type of technology we use most. The algorithms are Apriori, GRI, and Carma. Among them, Apriori and Carma mainly improve the analysis efficiency of the association rules, while GRI focuses on how to generalize the association of a single concept level to a more conceptual level, thereby revealing the internal structure of things. There are two main data storage forms for simple association rules. One is the transaction data format, and the other is the tabular data format.

Before starting association analysis, we need to do discretization for continuous data. Discretization is used in data analysis, especially in data mining. Commonly used, the main reasons are: a) The algorithm needs. Some data mining algorithms cannot use continuous variables directly and must be discretized before they can be included in the calculation. b) Reduce the sensitivity of abnormal data and make the model more stable. c) It facilitates the diagnosis and description of nonlinear relationships: After discrete processing of continuous data, the relationship between independent variables and target variables becomes clear.[3]

In this project, we will still work on the U.S. Permanent Visa dataset, implement Apriori algorithms to see the association rules between applicants' nationality and other features and answer following questions:

- Is the finding could improve what we have assumed before?
- Is there any new interesting finding?

- How could the association rules be used in practice?

III. Data Preparation

We choose the *SOC_NAME_short* (generate based on *pw_soc_title*), *citizenship* (generate based on *country_of_citizenship* and *country_of_citzenship*), *class_of_ad* (generate based on *class_of_admission*), *state* (generate based on *job_info_work_state*), *case_status*, *pw_amount* (generate based on *pw_amount_9089*) to start association analysis between different nationality records. We will focus on 4 Asian countries' records this time: China, Japan, South Korea and Philippines, because we believe that there is something interesting between these four countries based on the clustering results.

a. Each data is cleaned in project 3.

Thus, the whole dataset used for association analyzing is containing *case_status*, *SOC_NAME_short*, *citizenship*, *pw_amount*, *class_of_ad*, and *state*.

For implementing Apriopri algorithm, we need to discretization the continuous data *pw_amount*.

Two key problems in association with discretization are how to select the number of intervals or bins and how to decide on their width. Discretization can be performed with or without taking class information, if available, into account. These are the supervised and unsupervised ways. If class labels were known in the training data, the discretization method ought to take advantage of it, especially if the subsequently used learning algorithm for model building is supervised. In this situation, a discretization method should maximize the interdependence between the variable values and the class labels. Due to the existing of extremum, using *interval* or *cluster* methods will lead to an unbalance allocation, which will make the wage become useless in the analysis. Thus, we choose the *frequency* method to discretization pw_amount. Also, owe to there is only one feature need to be discretized in our dataset, we do not need to use different discretization methods.

The results of discretization are shown as following:

i) Subset1--China

```
          case_status                pw_amount_9089  class_of_admission    job_info_work_state
Certified        :13491   [7.7,7.24e+04)     :8808   H-1B       :21374    CALIFORNIA    :9840
Certified-Expired:10609   [7.24e+04,9.69e+04):8785   L-1        : 1837    NEW YORK      :2709
Denied           : 1076   [9.69e+04,1.35e+07]:8842   F-1        : 1515    WASHINGTON    :2334
Withdrawn        : 1259                              Not in USA: 1401    TEXAS         :1585
                                                     F-2        :   41    MASSACHUSETTS:1009
                                                     L-2        :   40    ILLINOIS      : 974
                                                     (Other)    :  227    (Other)       :7984

SOC_NAME_short
Softwa :9835
Electr :2293
Comput :1615
Statis :1427
Accoun :1212
Financ :1144
(Other):8909
```

ii) Subset2—South Korea

```
           case_status              pw_amount_9089  class_of_admission     job_info_work_state
Certified         :8516   [6.14,3.58e+04)    :6109   H-1B       :6487     CALIFORNIA     :5717
Certified-Expired:6602   [3.58e+04,6.31e+04):6160   F-1        :4404     NEW YORK       :1781
Denied            :1664   [6.31e+04,1.15e+07]:6137   Not in USA:2653     GEORGIA        :1342
Withdrawn         :1624                              E-2        :2124     TEXAS          :1258
                                                     B-2        : 863     NORTH CAROLINA:1254
                                                     F-2        : 645     NEW JERSEY     : 950
                                                     (Other)    :1230     (Other)        :6104

SOC_NAME_short
Meat,   : 1387
Softwa :  902
Market :  863
Accoun :  782
Electr :  714
Comput :  557
(Other):13201
```

iii) Subset3—Japan

```
           case_status              pw_amount_9089  class_of_admission job_info_work_state
Certified         :1321   [10,5.43e+04)      :911   H-1B       :1690     CALIFORNIA:1155
Certified-Expired:1052   [5.43e+04,8.59e+04):911   E-2        : 340     NEW YORK  : 464
Denied            : 239   [8.59e+04,2.27e+05]:912   F-1        : 248     WASHINGTON: 104
Withdrawn         : 122                             L-1        : 231     ILLINOIS  :  96
                                                    E-1        :  76     MICHIGAN  :  84
                                                    Not in USA:  48     TEXAS     :  84
                                                    (Other)    : 101     (Other)   : 747

SOC_NAME_short
Market : 307
Softwa : 227
Accoun : 212
Comput : 136
Financ :  89
Electr :  86
(Other):1677
```

iv) Subset4—Philippines

```
           case_status              pw_amount_9089  class_of_admission job_info_work_state
Certified         :3241   [7.55,4.25e+04)    :2581   H-1B       :5594     CALIFORNIA:1895
Certified-Expired:2863   [4.25e+04,6.95e+04):2575   B-2        : 666     TEXAS     : 694
Denied            :1220   [6.95e+04,5.56e+06]:2587   Not in USA: 636     NEW YORK  : 676
Withdrawn         : 419                              L-1        : 211     MARYLAND  : 479
                                                     F-1        : 143     ARIZONA   : 321
                                                     H-2B       : 131     NEW MEXICO: 317
                                                     (Other)    : 362     (Other)   :3361

SOC_NAME_short
Medica : 795
Home H : 710
Specia : 577
Occupa : 505
Softwa : 399
Second : 398
(Other):4359
```

IV. Modeling

First, we have found out the frequent itemset.

1. Frequent Itemset

The frequent itemset can be seen as collections of items that often appear together.[4] Here we look at the 10 most frequent items (relative frequency (=support) of items) in the data set. Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations).

The set of possible itemsets is the power set over I and has size $2^n - 1$ (excluding the empty set which is not a valid itemset). Although the size of the power-set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support [5] (also called anti-monotonicity[6]) which guarantees that for a frequent itemset, all its subsets are also frequent and thus no infrequent itemset can be a subset of a frequent itemset. Additionally, Exploiting this property, efficient algorithms (e.g., *Apriori*) can find all frequent itemsets.[7]

we use *itemFrequencyPlot()* function to show the frequent itemset (Figure 4.1-4.4):

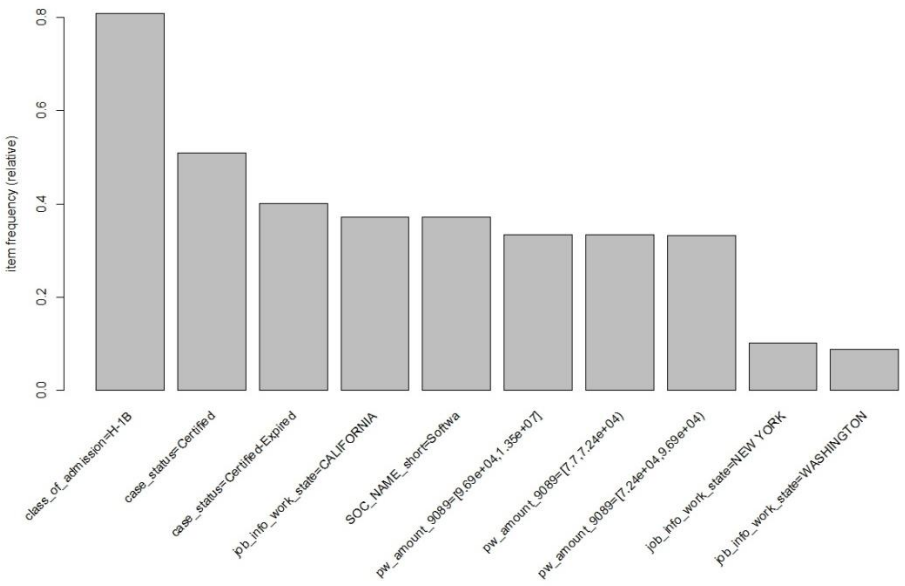

Figure 4.1(a) Subset1—China

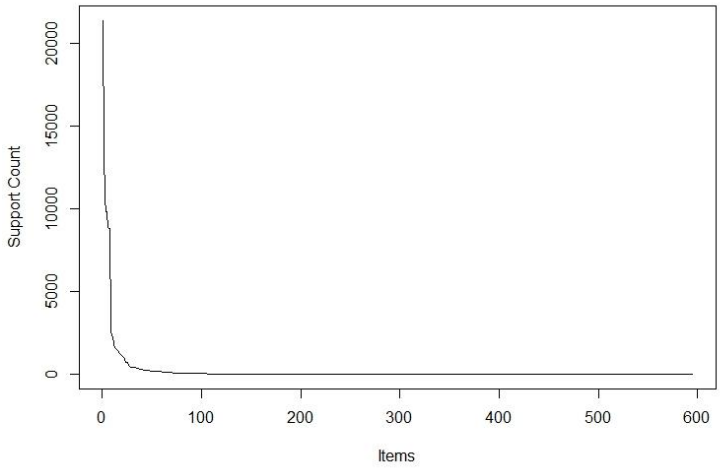

Figure 4.1(b) Subset1—China

```
[474] {case_status=Certified,
       country_of_citizenship=INDIA,
       pw_amount_9089=[9.87e+04,1.35e+07],
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Comput}                  0.0067492818  2361
[475] {case_status=Certified-Expired,
       country_of_citizenship=INDIA,
       pw_amount_9089=[7.57e+04,9.87e+04),
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Softwa}                  0.0064348298  2251
[476] {case_status=Certified,
       country_of_citizenship=INDIA,
       pw_amount_9089=[7.57e+04,9.87e+04),
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Softwa}                  0.0070351471  2461
[477] {case_status=Certified-Expired,
       country_of_citizenship=INDIA,
       pw_amount_9089=[6.14,7.57e+04),
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Softwa}                  0.0007232394   253
[478] {case_status=Certified,
       country_of_citizenship=INDIA,
       pw_amount_9089=[6.14,7.57e+04),
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Softwa}                  0.0005345683   187
[479] {case_status=Certified-Expired,
       country_of_citizenship=INDIA,
       pw_amount_9089=[9.87e+04,1.35e+07],
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Softwa}                  0.0190414934  6661
[480] {case_status=Certified,
       country_of_citizenship=INDIA,
       pw_amount_9089=[9.87e+04,1.35e+07],
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Softwa}                  0.0233323328  8162
```

Part of the frequent itemset

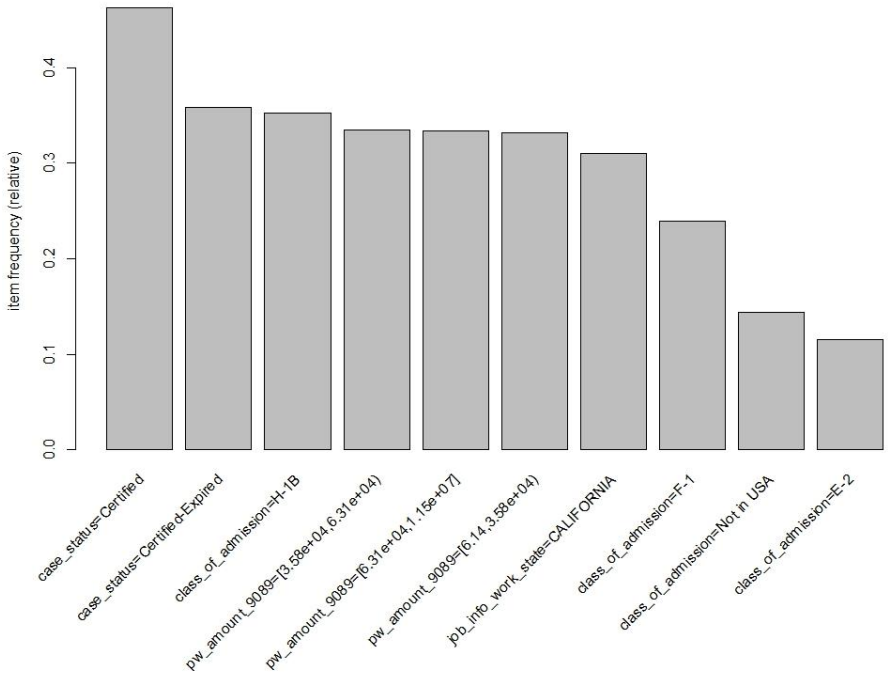

Figure 4.2(a) Subset2—South Korea

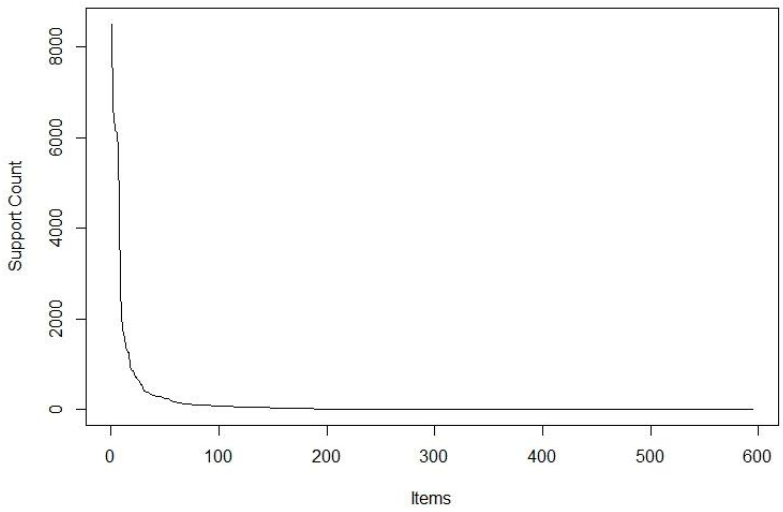

Figure 4.2(b) Subset2—South Korea

```
[578] {case_status=Denied,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=F-1,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0011409323    21
[579] {case_status=Certified-Expired,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=E-2,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0003259807     6
[580] {case_status=Certified,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=E-2,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0005976312    11
[581] {case_status=Certified-Expired,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=Not in USA,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0035857872    66
[582] {case_status=Certified,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=Not in USA,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0056503314   104
[583] {case_status=Certified-Expired,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=F-1,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0010866022    20
[584] {case_status=Certified,
       pw_amount_9089=[6.14,3.58e+04),
       class_of_admission=F-1,
       job_info_work_state=GEORGIA,
       SOC_NAME_short=Meat, }                    0.0026621754    49
```

Part of the frequent itemset

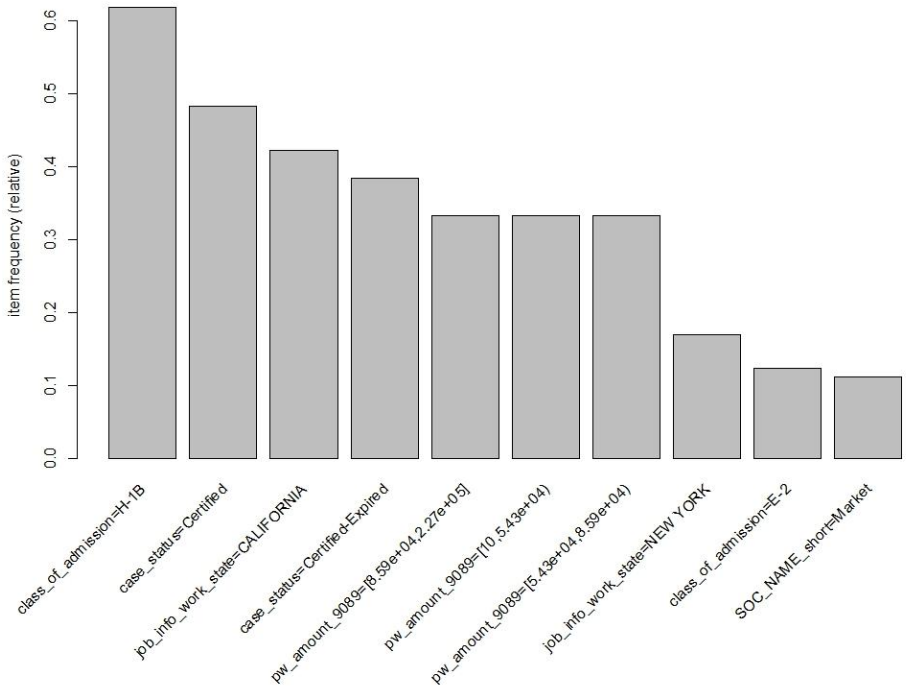

Figure 4.3(a) Subset3—Japan

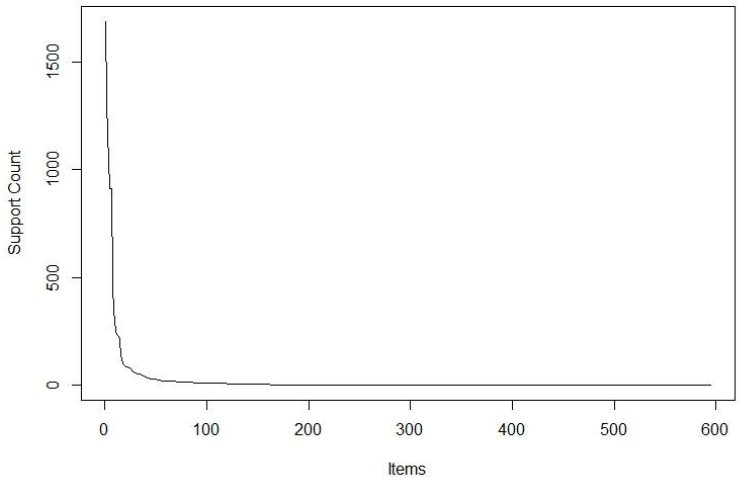

Figure 4.3(b) Subset3—Japan

	items	support	count
[1]	{case_status=Certified, pw_amount_9089=[10,5.43e+04), class_of_admission=E-2, job_info_work_state=CALIFORNIA, SOC_NAME_short=Food S}	0.001828822	5
[2]	{case_status=Certified, pw_amount_9089=[8.59e+04,2.27e+05], class_of_admission=H-1B, job_info_work_state=CALIFORNIA, SOC_NAME_short=Materi}	0.003291880	9
[3]	{case_status=Certified, pw_amount_9089=[10,5.43e+04), class_of_admission=F-1, job_info_work_state=CALIFORNIA, SOC_NAME_short=Interp}	0.001828822	5
[4]	{case_status=Certified, pw_amount_9089=[10,5.43e+04), class_of_admission=H-1B, job_info_work_state=CALIFORNIA, SOC_NAME_short=Interp}	0.001828822	5
[5]	{case_status=Certified, pw_amount_9089=[10,5.43e+04), class_of_admission=F-1, job_info_work_state=CALIFORNIA, SOC_NAME_short=Chefs }	0.001828822	5
[6]	{case_status=Certified, pw_amount_9089=[10,5.43e+04), class_of_admission=E-2, job_info_work_state=CALIFORNIA, SOC_NAME_short=Chefs }	0.002194587	6
[7]	{case_status=Denied, pw_amount_9089=[10,5.43e+04), class_of_admission=F-1, job_info_work_state=CALIFORNIA, SOC_NAME_short=Hairdr}	0.002926116	8
[8]	{case_status=Certified-Expired, pw_amount_9089=[10,5.43e+04), class_of_admission=F-1, job_info_work_state=CALIFORNIA, SOC_NAME_short=Hairdr}	0.004023409	11

Part of frequent itemset

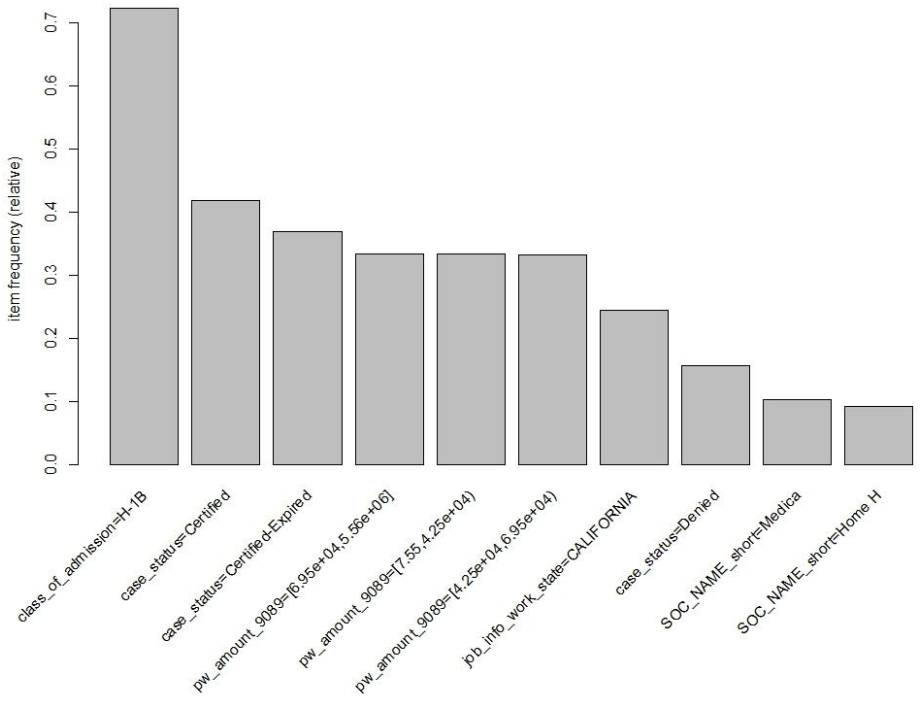

Figure 4.4(a) Subset4—Philippines

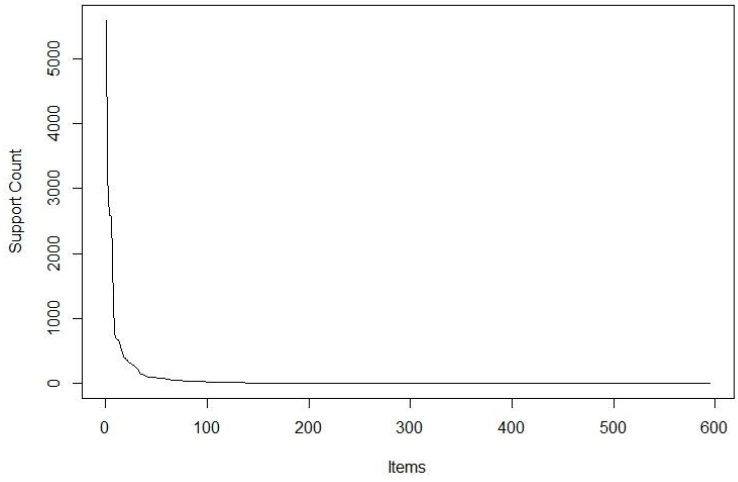

Figure 4.4(b) Subset4—Philippines

```
[330] {case_status=Certified,
       pw_amount_9089=[4.25e+04,6.95e+04),
       class_of_admission=H-1B,
       job_info_work_state=TEXAS,
       SOC_NAME_short=Medica}                    0.0036161694    28
[331] {case_status=Certified-Expired,
       pw_amount_9089=[7.55,4.25e+04),
       class_of_admission=H-1B,
       job_info_work_state=TEXAS,
       SOC_NAME_short=Medica}                    0.0018080847    14
[332] {case_status=Certified-Expired,
       pw_amount_9089=[4.25e+04,6.95e+04),
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Medica}                    0.0020663825    16
[333] {case_status=Certified,
       pw_amount_9089=[4.25e+04,6.95e+04),
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Medica}                    0.0027121271    21
[334] {case_status=Certified-Expired,
       pw_amount_9089=[6.95e+04,5.56e+06],
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Medica}                    0.0018080847    14
[335] {case_status=Certified,
       pw_amount_9089=[6.95e+04,5.56e+06],
       class_of_admission=H-1B,
       job_info_work_state=CALIFORNIA,
       SOC_NAME_short=Medica}                    0.0038744673    30
```

Part of frequent itemset

From the figures above (Figure 4.1(a)-4.4(a)), it is easily to find that those frequent itemsets prove some assumes we supposed before, such as Chinese applicants most work as IT engineers while Filipinos are most work in service industry.

2. Find a Reasonable Low Support

How to set reasonable support and confidence?

"For a rule: (A=a)->(B=b) (support=30%, confident=60%); where support=30% means that A=a and B= appear in all data records at the same time. The probability of b is 30%; confidence=60% means that in all data records, the probability of occurrence of B=b in the case of A=a is 60%, which is the conditional probability. The degree of support reveals the probability that A=a and B=b appears simultaneously. The degree of confidence reveals the probability that B=b will occur when A=a appears.

(1) If the support and confidence settings are too high, although the mining time can be reduced, it is easy to cause some infrequent feature items hidden in the data to be ignored, and it is difficult to find enough useful rules;

(2) If the support and confidence settings are too low, there may be too many rules, and even a lot of redundant and invalid rules. At the same time, due to the inherent problems of the algorithm, it will lead to high load calculations. Greatly increase the time of excavation."[4]

Here we use *5/nrow()* to calculate the support.

The results of calculating of each case are showing as Table 4.1. Because the calculating results may not as narrow as we hope, so we have tried to narrow it by ourselves. They are also showing the final low support we chosen to use in the table.

Table 4.1

	Calculated	**Final used**	**Confidence**
China	0.0001891432	0.02	0.9
South Korea	0.0002716505	0.00035	0.9
Japan	0.001828822	0.008	0.9
Philippines	0.0006457445	0.03	0.9

Now, let's plot them out (Figure 4.5-4.8).

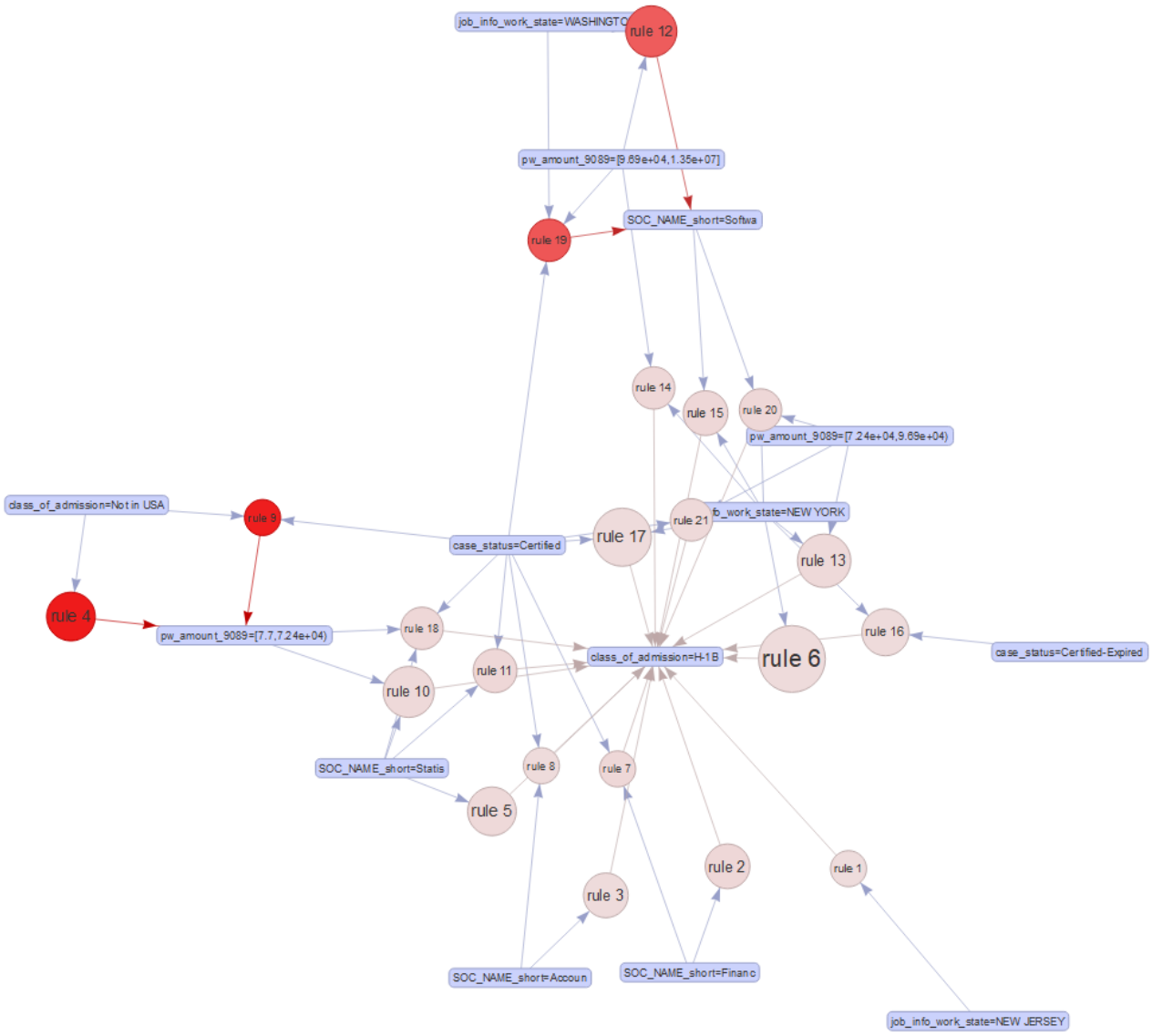

Figure 4.5 Subset1—China

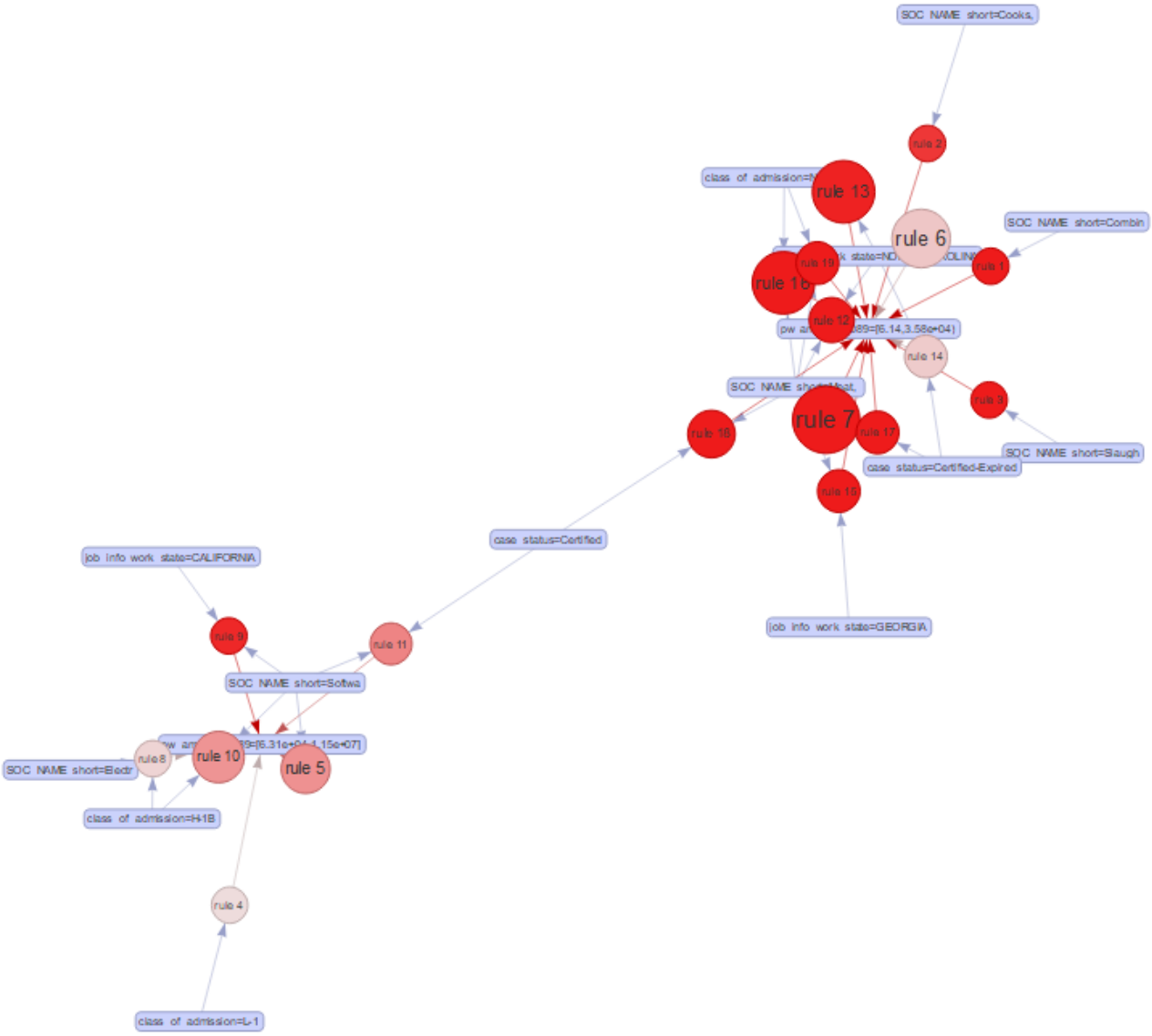

Figure 4.6 Subset2—South Korea

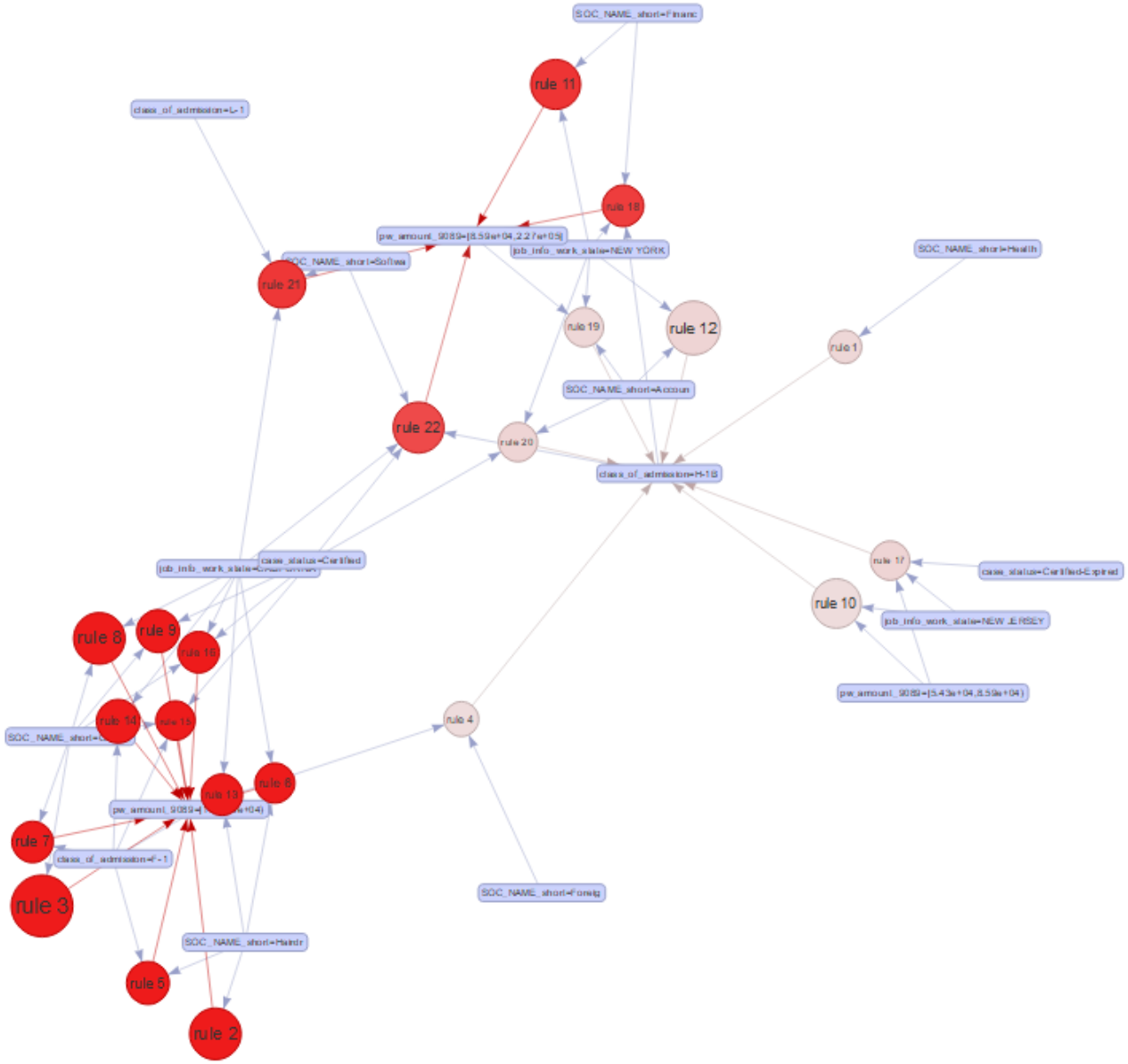

Figure 4.7 Subset3—Japan

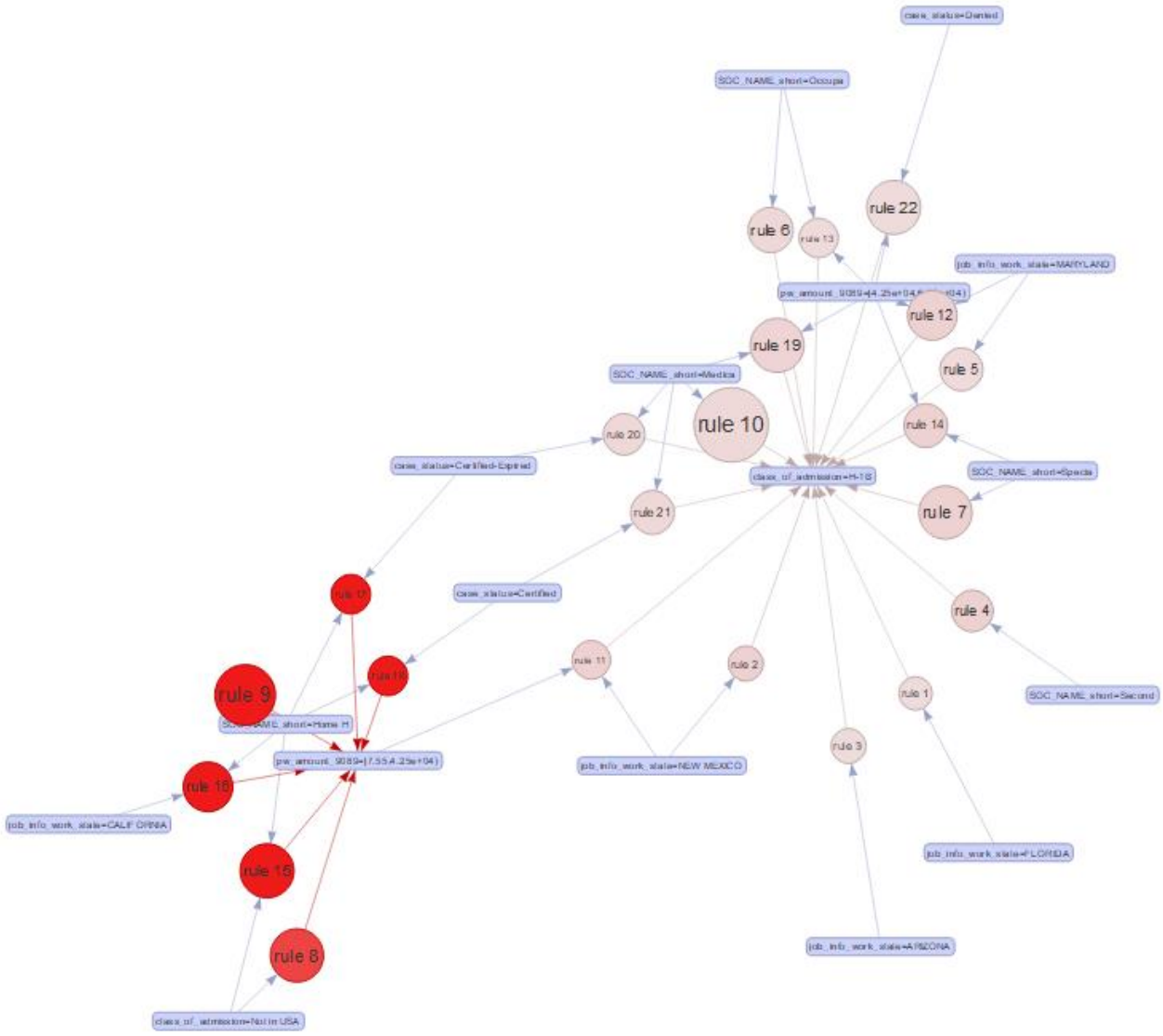

Figure 4.8 Subset4—Philippines

Then, because our main target is to see the principal factors that influence certified/ denied rate, so we improve our code to filtering the rules.

Factors of Denied Rate

First, we are focus on denied. we tabled the rules to see it clearly (Figure 4.5-4.8).

i) China (sorted by lift)

Extra analysis: The lift of the rule is actually the ratio of confidence to the subsequent term. The degree of improvement reflects the degree of influence of commodity A on commodity B. Greater than 1 is meaningful, that is, the promotion of A for B, the bigger the better.

```
    lhs                                   rhs                   support  confidence     lift  count
[1] {job_info_work_state=OHIO,
     SOC_NAME_short=Packer}            => {case_status=Denied} 0.0003782864   0.5882353 14.45167    10
[2] {job_info_work_state=FLORIDA,
     SOC_NAME_short=Cooks,}            => {case_status=Denied} 0.0003782864   0.5882353 14.45167    10
[3] {class_of_admission=Not in USA,
     job_info_work_state=OHIO,
     SOC_NAME_short=Packer}            => {case_status=Denied} 0.0003782864   0.5882353 14.45167    10
[4] {pw_amount_9089=[7.7,7.24e+04),
     job_info_work_state=OHIO,
     SOC_NAME_short=Packer}            => {case_status=Denied} 0.0003782864   0.5882353 14.45167    10
[5] {class_of_admission=Not in USA,
     job_info_work_state=FLORIDA,
     SOC_NAME_short=Cooks,}            => {case_status=Denied} 0.0003782864   0.5882353 14.45167    10
[6] {pw_amount_9089=[7.7,7.24e+04),
     job_info_work_state=FLORIDA,
     SOC_NAME_short=Cooks,}            => {case_status=Denied} 0.0003782864   0.5882353 14.45167    10
```

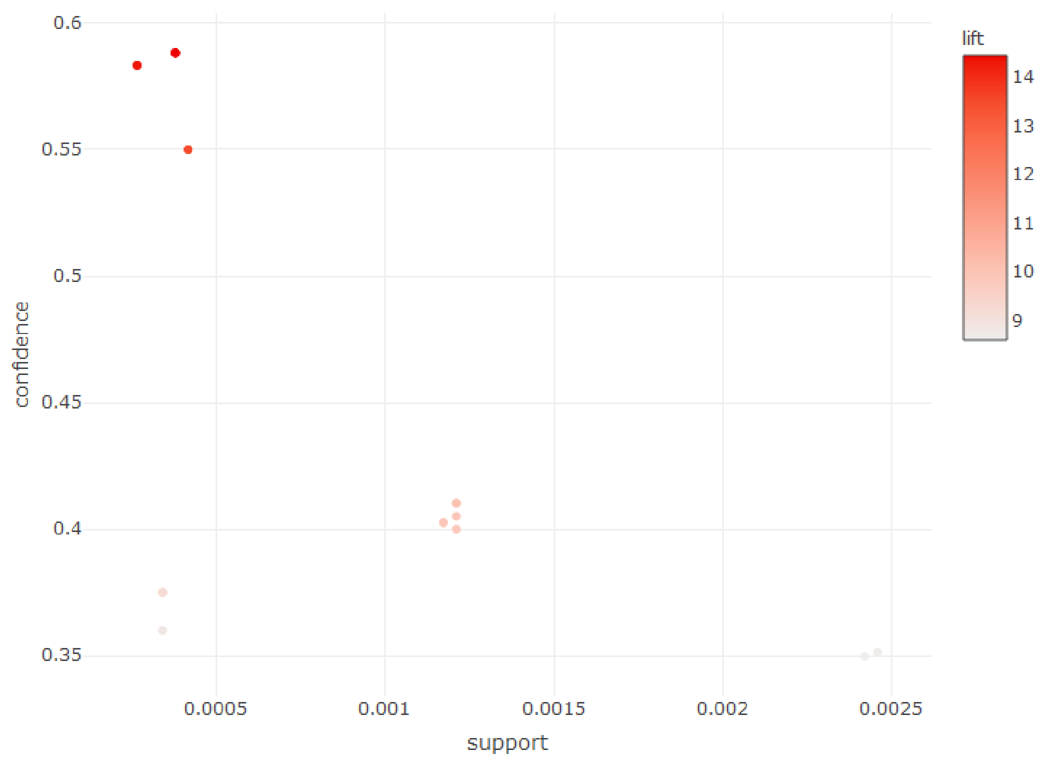

Figure 4.5 Subset1—China

From the top 6 sorted rules of Chinese applicants' records, we can see that those who work in ***Florida*** as a ***Cooker***, wage is in ***[7.7, 7.24e+04)***, ***not in the USA*** when applying, are more likely to be denied (rule 2,5,6); he who work in ***Ohio*** as a ***Packer***, wage is in ***[7.7, 7.24e+04)***, ***not in the USA*** when applying, are more likely to be denied, too. It is a useful guideline for Chinese applicants and their employers.

ii) South Korea (table ordered by support desc)

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[2]	{SOC_NAME_short=Admini}	{case_status=Denied}	0.001	0.306	3.380	11.000
[14]	{pw_amount_9089=[6.31e+04,1.15e+07],class_of_admission=F-1,job_info_work_state=VIRGINIA}	{case_status=Denied}	0.001	0.324	3.579	11.000
[16]	{class_of_admission=E-2,job_info_work_state=GEORGIA,SOC_NAME_short=Meat, }	{case_status=Denied}	0.001	0.333	3.687	11.000
[18]	{pw_amount_9089=[6.14,3.58e+04),class_of_admission=E-2,job_info_work_state=GEORGIA,SOC_NAME_short=Meat, }	{case_status=Denied}	0.001	0.333	3.687	11.000
[1]	{SOC_NAME_short=Second}	{case_status=Denied}	0.000	0.375	4.148	9.000
[7]	{job_info_work_state=TEXAS,SOC_NAME_short=Home H}	{case_status=Denied}	0.000	0.500	5.531	8.000
[9]	{class_of_admission=F-2,job_info_work_state=MARYLAND}	{case_status=Denied}	0.000	0.308	3.403	8.000
[13]	{pw_amount_9089=[6.14,3.58e+04),job_info_work_state=TEXAS,SOC_NAME_short=Home H}	{case_status=Denied}	0.000	0.500	5.531	8.000
[3]	{job_info_work_state=VIRGINIA,SOC_NAME_short=Second}	{case_status=Denied}	0.000	0.583	6.452	7.000
[4]	{pw_amount_9089=[6.31e+04,1.15e+07],SOC_NAME_short=Admini}	{case_status=Denied}	0.000	0.318	3.520	7.000

Showing 1 to 10 of 18 entries Previous 1 2 Next

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[5]	{class_of_admission=F-1,SOC_NAME_short=Foreig}	{case_status=Denied}	0.000	0.350	3.871	7.000
[6]	{job_info_work_state=CALIFORNIA,SOC_NAME_short=Foreig}	{case_status=Denied}	0.000	0.318	3.520	7.000
[8]	{class_of_admission=B-2,SOC_NAME_short=Sales }	{case_status=Denied}	0.000	0.368	4.075	7.000
[10]	{class_of_admission=F-1,job_info_work_state=CALIFORNIA,SOC_NAME_short=Foreig}	{case_status=Denied}	0.000	0.368	4.075	7.000
[11]	{pw_amount_9089=[3.58e+04,6.31e+04),class_of_admission=F-1,SOC_NAME_short=Foreig}	{case_status=Denied}	0.000	0.350	3.871	7.000
[12]	{pw_amount_9089=[3.58e+04,6.31e+04),job_info_work_state=CALIFORNIA,SOC_NAME_short=Foreig}	{case_status=Denied}	0.000	0.318	3.520	7.000
[15]	{pw_amount_9089=[3.58e+04,6.31e+04),class_of_admission=B-2,job_info_work_state=TEXAS}	{case_status=Denied}	0.000	0.304	3.366	7.000
[17]	{pw_amount_9089=[3.58e+04,6.31e+04),class_of_admission=F-1,job_info_work_state=CALIFORNIA,SOC_NAME_short=Foreig}	{case_status=Denied}	0.000	0.368	4.075	7.000

Showing 11 to 18 of 18 entries Previous 1 2 Next

Figure 4.6 Subset2—South Korea

By reading the table of rules of Korean applicants' records, it is obviously that those work in ***Virginia*** as a ***Secondary School Teachers*** are most likely to be denied; additionally, those work in ***Texas*** and SOC name is ***Home Health Aides***, wage is within ***[6.14, 3.58e+04)*** are more likely to be denied as well.

iii) Japan

```
    lhs                                                rhs                      support confidence      lift count
[1] {pw_amount_9089=[5.43e+04,8.59e+04),
     class_of_admission=VWT}                    => {case_status=Denied} 0.001463058  1.0000000 11.439331     4
[2] {class_of_admission=VWT,
     job_info_work_state=CALIFORNIA}            => {case_status=Denied} 0.001828822  1.0000000 11.439331     5
[3] {pw_amount_9089=[5.43e+04,8.59e+04),
     class_of_admission=VWT,
     job_info_work_state=CALIFORNIA}            => {case_status=Denied} 0.001463058  1.0000000 11.439331     4
[4] {class_of_admission=VWT}                    => {case_status=Denied} 0.001828822  0.8333333  9.532775     5
[5] {job_info_work_state=CALIFORNIA,
     SOC_NAME_short=Execut}                     => {case_status=Denied} 0.001463058  0.6666667  7.626220     4
[6] {class_of_admission=E-2,
     job_info_work_state=NEVADA}                => {case_status=Denied} 0.001463058  0.6666667  7.626220     4
```

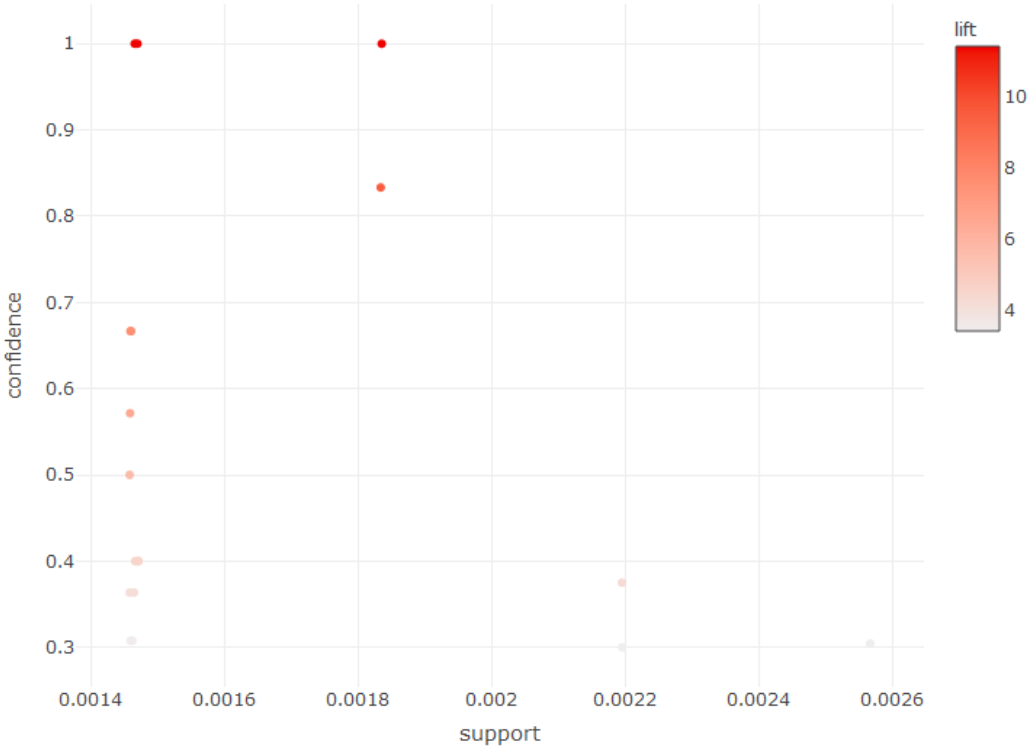

Figure 4.7 Subset3—Japan

From the Figures above, we can find out that for Japanese applicants, those who has a ***VWT*** state, work in ***California***, wage is in the ***[5.43e+04, 8.09e+04)***, are most likely to be denied.

iv) Philippines

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{SOC_NAME_short=Elemen}	{case_status=Denied}	0.010	0.369	2.343	79.000
[2]	{SOC_NAME_short=Middle}	{case_status=Denied}	0.011	0.361	2.290	83.000
[3]	{SOC_NAME_short=Second}	{case_status=Denied}	0.017	0.327	2.073	130.000
[4]	{job_info_work_state=MARYLAND}	{case_status=Denied}	0.025	0.409	2.597	196.000
[5]	{SOC_NAME_short=Specia}	{case_status=Denied}	0.025	0.334	2.123	193.000
[6]	{class_of_admission=H-1B,SOC_NAME_short=Middle}	{case_status=Denied}	0.011	0.361	2.290	83.000
[7]	{class_of_admission=H-1B,SOC_NAME_short=Second}	{case_status=Denied}	0.016	0.321	2.040	126.000
[8]	{job_info_work_state=MARYLAND,SOC_NAME_short=Specia}	{case_status=Denied}	0.011	0.477	3.026	82.000
[9]	{pw_amount_9089=[4.25e+04,6.95e+04),job_info_work_state=MARYLAND}	{case_status=Denied}	0.023	0.468	2.968	181.000
[10]	{class_of_admission=H-1B,job_info_work_state=MARYLAND}	{case_status=Denied}	0.024	0.419	2.660	184.000

Showing 1 to 10 of 17 entries Previous 1 2 Next

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[11]	{pw_amount_9089=[4.25e+04,6.95e+04),SOC_NAME_short=Specia}	{case_status=Denied}	0.016	0.383	2.428	127.000
[12]	{class_of_admission=H-1B,SOC_NAME_short=Specia}	{case_status=Denied}	0.025	0.337	2.141	193.000
[13]	{pw_amount_9089=[4.25e+04,6.95e+04),job_info_work_state=MARYLAND,SOC_NAME_short=Specia}	{case_status=Denied}	0.011	0.480	3.043	82.000
[14]	{class_of_admission=H-1B,job_info_work_state=MARYLAND,SOC_NAME_short=Specia}	{case_status=Denied}	0.011	0.477	3.026	82.000
[15]	{pw_amount_9089=[4.25e+04,6.95e+04),class_of_admission=H-1B,job_info_work_state=MARYLAND}	{case_status=Denied}	0.023	0.466	2.960	180.000
[16]	{pw_amount_9089=[4.25e+04,6.95e+04),class_of_admission=H-1B,SOC_NAME_short=Specia}	{case_status=Denied}	0.016	0.386	2.450	127.000
[17]	{pw_amount_9089=[4.25e+04,6.95e+04),class_of_admission=H-1B,job_info_work_state=MARYLAND,SOC_NAME_short=Specia}	{case_status=Denied}	0.011	0.480	3.043	82.000

Showing 11 to 17 of 17 entries Previous 1 2 Next

Figure 4.8 Subset4—Philippines

By looking at this table of rules, we can read that for Filipinos, those has a ***H-1B*** state, work in ***Maryland*** as a ***Special Education Teachers***, wage is within ***[4.25e+04, 6.95e+04)***, are more likely to be denied.

Factors of Certified Rate

i) China (table ordered by lift)

Then, we are focus on certified. we tabled the rules to see it clearly (Figure 4.9-4.12).

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[21]	{pw_amount_9089=[9.69e+04,1.35e+07],job_info_work_state=CALIFORNIA,SOC_NAME_short=Softwa}	{case_status=Certified}	0.089	0.590	1.157	2,345.000
[25]	{pw_amount_9089=[9.69e+04,1.35e+07],class_of_admission=H-1B,job_info_work_state=CALIFORNIA,SOC_NAME_short=Softwa}	{case_status=Certified}	0.073	0.590	1.156	1,918.000
[14]	{pw_amount_9089=[9.69e+04,1.35e+07],SOC_NAME_short=Softwa}	{case_status=Certified}	0.121	0.586	1.148	3,211.000
[22]	{pw_amount_9089=[9.69e+04,1.35e+07],class_of_admission=H-1B,SOC_NAME_short=Softwa}	{case_status=Certified}	0.094	0.584	1.145	2,486.000
[15]	{pw_amount_9089=[9.69e+04,1.35e+07],job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.122	0.571	1.119	3,238.000
[23]	{pw_amount_9089=[9.69e+04,1.35e+07],class_of_admission=H-1B,job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.101	0.571	1.119	2,671.000
[5]	{pw_amount_9089=[9.69e+04,1.35e+07]}	{case_status=Certified}	0.189	0.564	1.104	4,984.000
[16]	{pw_amount_9089=[9.69e+04,1.35e+07],class_of_admission=H-1B}	{case_status=Certified}	0.152	0.561	1.098	4,020.000
[24]	{class_of_admission=H-1B,job_info_work_state=CALIFORNIA,SOC_NAME_short=Softwa}	{case_status=Certified}	0.091	0.553	1.084	2,407.000

Showing 1 to 10 of 25 entries Previous 1 2 3 Next

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[9]	{class_of_admission=H-1B,job_info_work_state=NEW YORK}	{case_status=Certified}	0.051	0.546	1.071	1,350.000
[2]	{job_info_work_state=NEW YORK}	{case_status=Certified}	0.056	0.542	1.063	1,469.000
[6]	{SOC_NAME_short=Softwa}	{case_status=Certified}	0.198	0.532	1.043	5,234.000
[18]	{class_of_admission=H-1B,SOC_NAME_short=Softwa}	{case_status=Certified}	0.161	0.532	1.042	4,255.000
[19]	{class_of_admission=H-1B,job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.165	0.531	1.040	4,351.000
[7]	{job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.197	0.528	1.035	5,196.000
[8]	{class_of_admission=H-1B}	{case_status=Certified}	0.417	0.516	1.010	11,020.000
[1]	{}	{case_status=Certified}	0.510	0.510	1.000	13,491.000
[13]	{pw_amount_9089=[7.7,7.24e+04),class_of_admission=H-1B}	{case_status=Certified}	0.123	0.500	0.981	3,264.000

Showing 11 to 20 of 25 entries Previous 1 2 3 Next

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[4]	{pw_amount_9089=[7.7,7.24e+04)}	{case_status=Certified}	0.161	0.484	0.949	4,267.000
[3]	{pw_amount_9089=[7.24e+04,9.69e+04)}	{case_status=Certified}	0.160	0.483	0.946	4,240.000
[11]	{pw_amount_9089=[7.24e+04,9.69e+04),job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.051	0.467	0.915	1,343.000
[20]	{pw_amount_9089=[7.24e+04,9.69e+04),class_of_admission=H-1B,SOC_NAME_short=Softwa}	{case_status=Certified}	0.055	0.462	0.906	1,464.000
[10]	{pw_amount_9089=[7.24e+04,9.69e+04),SOC_NAME_short=Softwa}	{case_status=Certified}	0.063	0.454	0.889	1,675.000

Showing 21 to 25 of 25 entries Previous 1 2 3 Next

Figure 4.9 Subset1—China

From the table of rules for Chinese applicants, we can see that those who has ***H-1B*** state, work in ***California*** as a ***Software Engineer***, wage is in ***[9.69e+04, 1.35e+07)***, are most likely to be certified.

ii) South Korea

```
    lhs                                                                rhs                       support    confidence lift     count
[1] {pw_amount_9089=[6.14,3.58e+04),class_of_admission=F-1}         => {case_status=Certified} 0.04873411 0.4901639  1.059413  897
[2] {job_info_work_state=NEW YORK}                                  => {case_status=Certified} 0.04677822 0.4834363  1.044872  861
[3] {class_of_admission=F-1}                                        => {case_status=Certified} 0.11354993 0.4745686  1.025706 2090
[4] {pw_amount_9089=[6.31e+04,1.15e+07],class_of_admission=H-1B}    => {case_status=Certified} 0.09138324 0.4727375  1.021748 1682
[5] {pw_amount_9089=[3.58e+04,6.31e+04)}                            => {case_status=Certified} 0.15652505 0.4676948  1.010849 2881
[6] {pw_amount_9089=[6.31e+04,1.15e+07]}                            => {case_status=Certified} 0.15516679 0.4653740  1.005833 2856
```

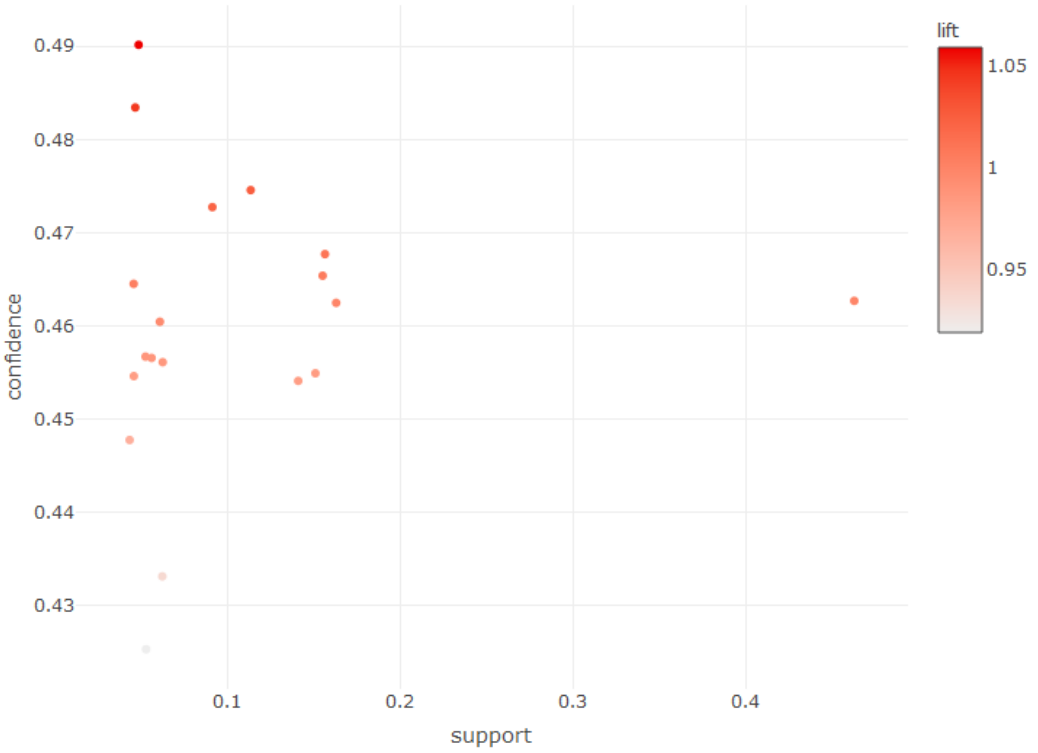

Figure 4.10 Subset2—South Korea

By reading the top 6 rules sorted by lift of Korean applicants' records, it is obviously that those work in ***New York*** with a ***F-1*** or ***H-1B*** state, wage is within ***[6.14, 3.58e+04)*** or is within ***[3.58e+04, 6.31e+04)*** or is within ***[6.31e+04, 1.15e+07]*** are most likely to be certified.

iii) Japan

```
    lhs                                                                rhs                          support    confidence lift     count
[1] {pw_amount_9089=[8.59e+04,2.27e+05],class_of_admission=L-1} => {case_status=Certified} 0.03291880 0.5389222  1.115377  90
[2] {pw_amount_9089=[8.59e+04,2.27e+05],SOC_NAME_short=Softwa} => {case_status=Certified} 0.03438186 0.5310734  1.099133  94
[3] {pw_amount_9089=[10,5.43e+04),class_of_admission=F-1}      => {case_status=Certified} 0.03547915 0.5271739  1.091062  97
[4] {class_of_admission=F-1,job_info_work_state=CALIFORNIA}    => {case_status=Certified} 0.03438186 0.5251397  1.086852  94
[5] {SOC_NAME_short=Softwa}                                    => {case_status=Certified} 0.04352597 0.5242291  1.084968 119
[6] {class_of_admission=F-1}                                   => {case_status=Certified} 0.04718361 0.5201613  1.076549 129
```

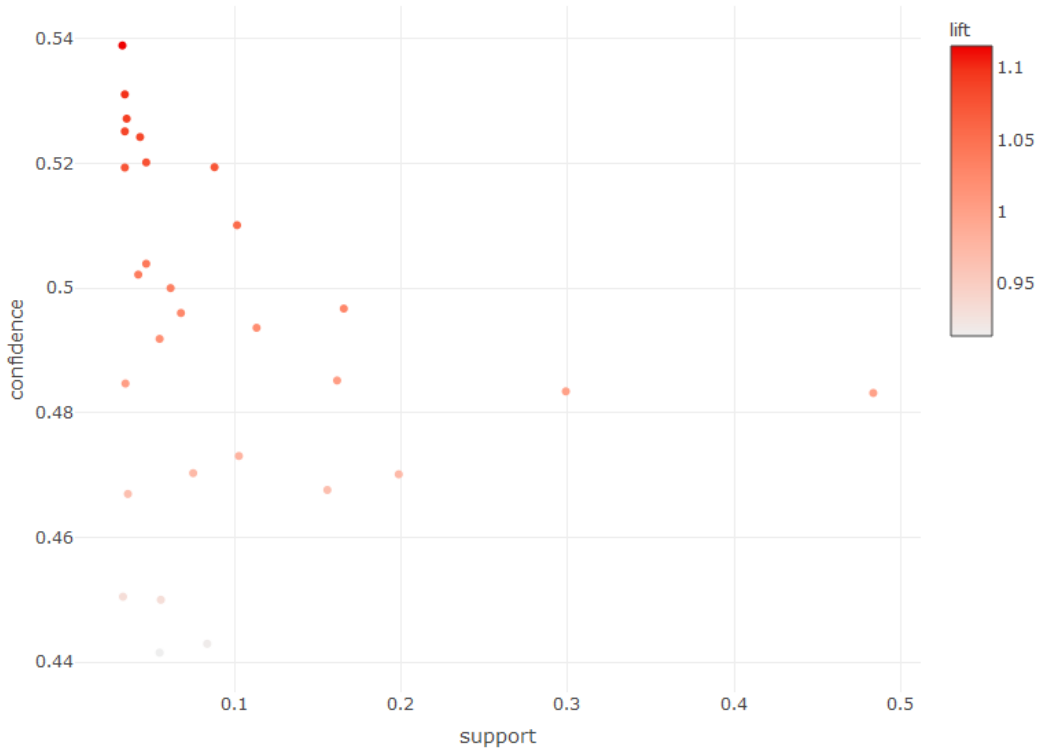

Figure 4.11 Subset3—Japan

From the Figures above, we can find out that for Japanese applicants, those who has a ***L-1*** or ***F-1*** state, work in ***California*** as a ***Software Engineer***, wage is in the ***[8.59e+04, 2.27e+05]***, are most likely to be certified.

iv) Philippines

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{}	{case_status=Certified}	0.419	0.419	1.000	3,241.000
[2]	{class_of_admission=Not in USA}	{case_status=Certified}	0.036	0.437	1.044	278.000
[3]	{class_of_admission=B-2}	{case_status=Certified}	0.031	0.362	0.865	241.000
[4]	{job_info_work_state=NEW YORK}	{case_status=Certified}	0.038	0.438	1.046	296.000
[5]	{job_info_work_state=TEXAS}	{case_status=Certified}	0.035	0.389	0.929	270.000
[6]	{SOC_NAME_short=Home H}	{case_status=Certified}	0.035	0.380	0.909	270.000
[7]	{SOC_NAME_short=Medica}	{case_status=Certified}	0.046	0.449	1.073	357.000
[8]	{job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.101	0.412	0.983	780.000
[9]	{pw_amount_9089=[4.25e+04,6.95e+04)}	{case_status=Certified}	0.133	0.399	0.953	1,027.000
[10]	{pw_amount_9089=[7.55,4.25e+04)}	{case_status=Certified}	0.125	0.374	0.894	966.000

Showing 1 to 10 of 21 entries Previous 1 2 3 Next

Show 25 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[11]	{pw_amount_9089=[6.95e+04,5.56e+06]}	{case_status=Certified}	0.161	0.482	1.153	1,248.000
[12]	{class_of_admission=H-1B}	{case_status=Certified}	0.304	0.421	1.005	2,353.000
[13]	{pw_amount_9089=[7.55,4.25e+04),class_of_admission=Not in USA}	{case_status=Certified}	0.031	0.409	0.978	237.000
[14]	{pw_amount_9089=[7.55,4.25e+04),SOC_NAME_short=Home H}	{case_status=Certified}	0.035	0.380	0.909	270.000
[15]	{class_of_admission=H-1B,SOC_NAME_short=Medica}	{case_status=Certified}	0.043	0.450	1.076	330.000
[16]	{pw_amount_9089=[7.55,4.25e+04),job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.033	0.369	0.881	256.000
[17]	{pw_amount_9089=[6.95e+04,5.56e+06],job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.046	0.453	1.081	353.000
[18]	{class_of_admission=H-1B,job_info_work_state=CALIFORNIA}	{case_status=Certified}	0.050	0.423	1.009	390.000
[19]	{pw_amount_9089=[4.25e+04,6.95e+04),class_of_admission=H-1B}	{case_status=Certified}	0.118	0.396	0.947	910.000
[20]	{pw_amount_9089=[7.55,4.25e+04),class_of_admission=H-1B}	{case_status=Certified}	0.058	0.365	0.871	452.000
[21]	{pw_amount_9089=[6.95e+04,5.56e+06],class_of_admission=H-1B}	{case_status=Certified}	0.128	0.482	1.150	991.000

Showing 1 to 21 of 21 entries Previous 1 Next

Figure 4.12 Subset4—Philippines

By looking at this table of rules, we can read that for Filipinos, those has a ***H-1B*** state, work in ***California*** or ***New York*** as a ***Medical related worker***, wage is within ***[6.95e+04, 5.56e+06]***, are more likely to be certified.

V. Evaluation

From the association analysis result of each subset, we can safely achieve the following conclusions:

- From these patterns, we can prove some assume we made before:
  a) ***Chinese*** applicants most work as ***IT engineers*** while ***Filipinos*** are most work in ***service industry***;
  b) ***Software Engineer*** is the hottest occupation for international workers and the USA employers;
  c) ***California***, ***Texas*** and ***New York*** are the hottest states within international workers;
  d) Applicants who owns a ***H-1B*** visa or ***F-1*** visa are more likely to be certified;
  e) Applicants who work as **Technique Staff** are more likely to own ***higher wage*** than those work in ***Service Industry***.

- ❖ Also, we can see some brand-new differences between different patterns (Figure 4.1-4.4):
  - a) Figure 4.1 shows that only 4 rules are strong rules, others are all have less influence;
  - b) Figure 4.2 shows that there are not only many strong rules, but also some medium influence rules. And they have tight relations;
  - c) Figure 4.3 shows that there are little weak rules, others are all strong ones;
  - d) Figure 4.4 shows that there are less strong rules, but the strong rules are having tight relations with each other while the weaker ones are having tight relation between their selves.

This difference may be caused by the different immigration policy for different country. For example, for the Chinese, they need to schedule before applying for a "Green Card"; on the opposed, there is no such limitation for the Korean. And there even goes some policies encourage the Korean to immigrate to the United States.

Extra analysis: The data shows that the peak of Japanese immigration to the United States was 1910-1920, 1950-1960, and there was very little immigration after 1970. Japanese-Americans began to decline in the 1980s, and the peak period of decline is in the 1990s. During 2000-2010, the number began to stabilize.[8]

Asian American alone [edit]

Ancestry	Population 2000	Population 2010	Percent change
Bangladeshi	46,905	142,080	202.9%
Bhutanese	192	18,814	9,699.0%
Burmese	14,620	95,536	553.5%
Cambodian	183,769	255,497	39.0%
Chinese	2,564,190	3,535,382	37.9%
Filipino	1,908,125	2,649,973	38.9%
Hispanic	119,829	–	–
Hmong	174,712	252,323	44.4%
Indian	1,718,778	2,918,807	69.8%
Indonesian	44,186	63,383	69.7%
Japanese	852,237	841,824	−1.2%
Korean	1,099,422	1,463,474	33.1%

Figure 4.13 Asian American Population[8]

On the other hand, we can also find some similarity between the 4 subsets, which may prove the similarity of all applicants:

- ***Packer***, ***Teacher***, ***Cooker*** are not the occupation which is needing many talents;
- ***Maryland***, ***Ohio***, ***Florida*** are not the suitable work state for applicants;

From the qualification requirements of the U.S. Permanent visa, it is easily to find that the international students and workers, the companies and organizations who want to hire international employees are the crucial stakeholders who want to know the rules between nationality and certification/ denied rate of the Green Card application: applicants are need to use the rules of certified to know which characteristic could help them to increase their competition; on the other side, they also need to use the denied rules to see which characteristic they need to omit for avoiding the risk of denied; the employers need to use the rules to help its employees to prepare their application details, which can reduce their workload—failed application means wasting time and increasing employment cost. Here are some recommendations:

- For applicants, check whether you wage is suitable for your occupation; check whether your occupation is the one that is needing many talents recently; check whether your work state is opening for international workers;
- For employees, check whether you international employer owns are suitable valid visa state; check your employer's nationality with the other tight related information.

Extra analysis: The International Migration Organization (IOM) and the Global Intelligence Group (CCG) jointly issued the "World Migration Report 2018". It points out that the number of international immigration is increasing. The migrant workers mainly engaged in the service industry have become the main body of international immigration. As of 2015, the United States is the largest destination country for immigration. In 2015, there were 46 million immigrants, and India is the world's largest exporter of immigrants, exporting more than 15 million immigrants.[9]

VI. Reference

[1] Association rule learning (https://en.wikipedia.org/wiki/Statistical_classification)

[2] Association analysis (https://blog.csdn.net/g090909/article/details/52785699)

[3] Discretization of Continuous Variables (http://www.ppvke.com/Blog/archives/44271)

[4] Machine Learning: Association Analysis (https://blog.csdn.net/kevinelstri/article/details/53487186)

[5] Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF). Introduction to Data Mining. Addison-Wesley. ISBN 0-321-32136-7.

[6] Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S.; Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, 2001, Heidelberg, Germany, 2001, pages 433-442

[7] Agrawal, Rakesh; and Srikant, Ramakrishnan; Fast algorithms for mining association rules in large databases, in Bocca, Jorge B.; Jarke, Matthias; and Zaniolo, Carlo; editors, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, September 1994, pages 487-499

[8] Demographics of Asian Americans (https://en.wikipedia.org/wiki/Demographics_of_Asian_Americans#Asian_American_alone)

[9] World Migration Report 2018 (https://www.iom.int/wmr/world-migration-report-2018)