Journal of Computers Vol. 31 No. 2, 2020, pp. 56-69
doi:10.3966/199115992020043102007

Mining Association Rules between Education, Family Background and Earning

Chih Yen Chang[1], Ming Sang Chang[2*]

[1] Graduate Institute of Communication Engineering, National Taiwan University, Taiwan, ROC
gsmmcc@gmail.com

[2] Department of Information Management, Central Police University, Taiwan, ROC
mschang@mail.cpu.edu.tw

Received 16 April 2018; Revised 11 January 2019; Accepted 9 February 2019

Abstract. Education is widely regarding as the primary way to allocate the economic remuneration. Most people consider education as a mechanism that can sabotage the association of economic inequality. In this paper, we intend to propose a mechanism to investigate the determinants that take influence on one's economic rewards by leveraging diverse data mining techniques, instead of using statistical methods solely. We found the determinants of one's earning would vary depending on macro or micro perspectives which are rarely mentioned in other works. In macro view, it showed that the self-worth of a student played the most important role, followed by school practice and family background. In micro view, the sub-factors were investigated with six scenarios respectively. Our findings not only reflect the status-quo of society, but can be references in how background shapes one's future.

Keywords: association rule mining, data mining, inequality, random forest similarity, socioeconomic status

1 Introduction

When concluding why a person would succeed, the factors are typically categorized into two types, innate and acquired. Innate factors are things can't be chosen in one's life, like family background and physical condition. Acquired factors are what you can choose to do, such as pursuing an education degree or leaning professional skills. However, success stories from political dynasties and bigwig families repeatedly remind us the dominance and importance of the innate factors. Most of the empirical studies suggested the existence of economic inequality transmitted across generations. Nonetheless, economists and sociologists have been disputed for the degree and mechanism of economic privilege actually transmitted from one generation to another. Family background, especially economic advantage, is served as a strong predictor of economic status in the next generation. Furthermore, some economists have pointed out that other factors may be important as well, such as education and non-cognitive behaviors.

Education is widely regarding as the primary way to allocate the economic remuneration. Basically, every study has implied the significant relationship between education achievement and personal earnings [1]. Examining the inheritance of economic privilege can be considered as a branch of issues that aims to find the determinants of one's life achievement. Achievement typically denotes as socioeconomic status (SES) which is a combined measure based on one's education, income and occupation in economic and sociological perspectives. The major theoretical view in economic literatures exploit the family behaviors by claiming families would transmit cultural and genetic characteristics to their offspring that effects children's SES [2]. Accordingly, a large number of empirical studies have discussed about relationship between family background and one's SES. Some other researches have

[*] Corresponding Author

investigated the existence and correlation of mediating factors between family background and one's SES [3]. One of the decisive mediating factors is educational attainment. Either the connection between family background and education attainment or between education attainment and adult's SES has been examined in many studies. As a premise, they normally treat these three facts with an ordinal manner, one's SES following education attainment and family background. Due to the restriction of data collected and archived, most of the researches were held in Western communities. Besides, statistical tools such as regression analysis are widely adopted in such studies. These tools perform well but still maintain some restrictions like assuming multivariate normality of independent variables [4].

In this paper, we propose a mechanism to investigate the determinants that take influence on one's economic rewards. In order to bypass the limitation mentioned, we leverage diverse data mining techniques, instead of using statistical methods solely. By utilizing association rule mining, the relationships can be easily interpreted with if-then rules. Drawing on data from Taiwan Education Panel Survey (TEPS) [5], one's background in the study stage can be represented completely in aspects of family, education and self-worth. TEPS and its continued survey tracked the economic and education outcomes for students who enrolled in the eleventh grade in 2001 and were periodically surveyed until 2009 with most aged 24 to 25 [6]. This study tries to break down into a subtler level to investigate sub-factors inside the examined determinants. The proposed method addresses results by what background conditions would lead to high or low economic rewards in future with highly supportive evidences.

In this study, our main contribution is that we found the determinants of one's earning would vary depending on macro or micro perspectives which are rarely mentioned in other works. In macro view, the relationship was discussed between earning and overall background of family, school and self-worth respectively. It showed that the self-worth of a student played the most important role, followed by school practice and family background. In micro view, the sub-factors were investigated in high and low earning groups separately. Most rules in high earing group are related with parent's working status while, in low earning group, they are frequently linked with student's self-worth.

The resting of this paper is organized as follows. Section 2 surveys data mining techniques. The system architecture of proposed method is described in section 3. In section 4, an actual dataset is applied as a case study to validate the capability of proposed mechanism. The results of the case study would be analyzed in section 5. Finally, conclusions are presented in section 6.

2 Related Works

Multivariate techniques are used to study data sets in social science research. These techniques are particularly important in social science research because social researchers are generally unable to use randomized laboratory experiments, like those used in natural sciences. Multivariate techniques can statistically estimate relationships between different variables, and correlate how important each one is to the final outcome and where dependencies exist between them. For multivariate techniques to give meaningful results, they need a large sample of data; otherwise, the results are meaningless due to high standard errors. Standard errors determine how confident you can be in the results, and you can be more confident in the results from a large sample than a small one. These tools may perform well but still maintain some restrictions like assuming multivariate normality of independent variables. In our paper, we want to bypass the limitation mentioned. We leverage diverse data mining techniques, instead of using statistical methods solely.

This section will briefly introduce the basic theoretical background about our study. We also draw attention to the strengths and limitations of each method through analysis.

In order to find the optimal tree, the goal is normally set to minimize the generalization error. Nonetheless, other goals or target functions can be alternatives as well, such as to minimize the average tree depth or minimize the number of nodes. We focus on Classification and Regression Tree (CART) [7]. It's featured as a binary tree which means there are exact two outgoing edges in each internal node. This restriction ensures there won't exist too many subsets after splitting which may lead to over-fitting.

Random Forest (RF) algorithm is efficient and powerful in solving classification and regression problems. An individual tree in RF is fully-grown and unpruned. It indicates that the terminal nodes will only accommodate a small amount of observations. As a feature selection mechanism, Boruta algorithm [8] aims to find all the relevant attributes. It's developed as a wrapper based on Random Forest algorithm.

There is a variety of types of clustering algorithm. They can be divided into two types, hierarchical and partitional. Constructing partitions with a hierarchical form, hierarchical clustering algorithm presents the partitions by a dendrogram that each group is the union of its sub-groups. The root group containing all the observations. On the other hand, partitional clustering algorithms simply divide observations into non-overlapping groups which mean each observation can only be in one group.

Clustering analysis intents to deal with partitioning a given dataset embedded in n-dimensional space into *k* different clusters. One of partitional clustering algorithms is the K-means algorithm [9]. It builds a prototype with a centroid, typically regarded as the mean of a group of points. Genetic K-means (GKA) [10] has been proposed and proven to converge efficiently to the global optimum in the clustering issue. It takes advantage of genetic algorithm to find the global minimum.

The correctness of deciding the number of clusters (*K*) is considered to be unclear. Choosing optimal *K* value is about to maintain a balance between accuracy of assigning each data to a correct cluster, and compression of data with clusters. There are several studies proposing *K* selection mechanisms each with different pros and cons. In this study, we adopt three classical methods simultaneously to solve this issue, which are elbow method [11], gap statistic [12] and prediction strength [13].

There are some major association mining techniques and algorithms. Apriori algorithm can be more efficient during the candidate generating process [14]. It avoids measuring some item sets by utilizing pruning techniques and can keep item sets not to be too large.

Normally, association rule mining algorithms generate an overwhelming number of rules due to the typically highly correlated transaction datasets [15]. There are several pruning methods that have been adopted to eliminate redundant rules and reduce number of rules kept [16]. Different pruning techniques are used in different steps of association mining process. Chi square testing is a hypothesis testing method for discrete type of data from statistics. It can evaluate the correlation between two variables and hence determine whether they are correlated or independent [17]. In our study, we take advantages of chi square testing in a rule pruning step. Chi-squared measure is utilized in many related studies, including lately work like [18], to rank produced rules.

3 System Architecture and Methodology

The goal of this study is to derive important insights from the given dataset. Rule-based representation of insights is chosen because of its easily understandable interpretation. In this study, we proposed 3 stages for deriving rule-based insights: clustering stage, rule generating stage, and rule extracting stage. In the clustering stage, clustering mechanism is utilized within the dataset before generating association rules. With clustering process, each group is expected to maintain associations without the interference from other clusters that help stand out the originally infrequent items in a specific group. On the other hand, after rules brought out in each group, there is still very likely to have tremendous amount of outcomes. In favor of finding out which rules are comparably important, rule pruning and rule ranking techniques are necessary in the rule extracting stage.

3.1 Clustering Stage

In clustering stage, the first task is to calculate the dissimilarity between each observation for the purpose of defining distance in the coordinate space. We use Random Forest algorithm to determine a supervised dissimilarity measure by taking advantages of its ability to deal with mixed types of data. As the class labels set, RF produces a similarity matrix between each object after classification, and the dissimilarity is defined as sqrt (1-similarity). The precision of dissimilarity measure is correlated with the accuracy of RF classification. It's shown that applying feature selection before the classification can decrease the out-of-bag error rate. To that end, Boruta algorithm [8] is chosen to select out the important features. By settling the rejection region to test variable importance value, Boruta algorithm helps identify the significant features that take impacts on classification results.

In case of maintaining various databases with multiple types of features, a combined dissimilarity matrix is constituted by aggregating dissimilarity measures from each distinct database. The combined matrix takes the influences of different features from multiple databases into account at the same time. A weighting value ranging from 0 to 1 is assigned to each dissimilarity matrix with the summation of these values to be 1. Grid search algorithm is utilized to determine the weighting values that give the minimal total within cluster variation (TWCV) value of the clustering outcome. Grid search is a full scale searching through a subset of the parameter space of a learning algorithm. As a result, it returns the set of parameters that keeps minimal TWCV value.

The clustering outcome for computing TWCV is derived by genetic k-means algorithm (GKA) [10]. GKA conducts clustering with the guarantee to converge at the global optimum of TWCV. The K value, number of cluster centroids, used in GKA is determined by three measurements: TWCV, Gap statistics and prediction strength value. All these measurements are calculated from K=1 to n, and K=p is chosen by elbow method that picks the point before the marginal gain dropping suddenly.

3.2 Rule Generation Stage

In rule generation stage, Apriori algorithm [14] takes responsibility to develop association rules. Considering the rule pruning process followed on, the selection of confidence and support thresholds is quite flexible and less sensitive to the system [16, 19]. Association rules are developed from each cluster respectively. We keep the amount of rules generated within a hecto-scale and treat them to further process in the next stage. In addition, the k-fold cross validation is utilized when generating rules in order to avoid over-fitting. All the observations in each cluster would be divided into k sets with equal size. The rules are hence produced within each set and examined among each other. The performance metrics are averaged with the given rule applied in all the (k-1) sets.

3.3 Rule Extraction Stage

In rule extracting stage, after rules developed within each cluster, the rule pruning techniques are applied to cut out redundant or insignificant rules. We choose redundant rule pruning to eliminate superfluous rules and Chi-squared test pruning to pick rules with positive correlation between the antecedent and the consequent [18]. Subsequently, the importance of rules is presented by rule ranks. For example, given the rules:

R_1: A => Y, with support = S_1, confidence = C_1 and chi-squared measure = X_1.

R_2: {A, B} => Y, with support = S_2, confidence = C_2 and chi-squared measure = X_2.

A and B are items in the shared item set with Y being the class label and consequent. It’s evident that the antecedent of R_2 is a subset of R_1’s antecedent. For both R_1 and R_2, rules would be kept if their chi-squared measure X_1 and X_2 are greater than the threshold X_t. X_t is a pre-defined threshold of chi-squared measure. Typically, with p-value = 0.05 and degree of freedom = 1, the threshold is set that X_t = 3.84. Furthermore, for pruning the redundancy, R_2 would be eliminated if the confidence C_2 is less than C_1. The combination of chi-squared pruning and redundant rule pruning help drastically reduce the number of rules mined and leave only the important rules qualified by the metrics.

Rules are given ranks with the criterions including confidence, support, lift, chi-squared measure and cardinality. Cardinality is regarded as the number of antecedents. Larger cardinality means to maintain more items as the antecedent that results in more specific rules. Chi-squared measure is first applied in sorting and then followed by lift, confidence, support and cardinality. To note that, each of the metric representing a rule is the average value from k sets via k-fold cross validation.

The ranking process is performed with a starting metric, and if the tie-break happened in the given metric, the priority is determined by the ranking principle. For example, for picking chi-squared measure as the starting metric, the tie-break occurred if two rules shared the same chi-squared measure. To solve the condition, the rule with higher lift value would be considered to have higher rank. Each of chi-squared measure, lift, support and confidence would be chosen to be the starting one and ranked by the mentioned principle. Hence, four different set of ranked rules would be stated correspondingly. With the aid of ranking process, we can extract rules which are relatively more important than others.

The system architecture is summarized in Fig. 1. Each process in the entire system is connected sequentially from top to bottom. Our system allows multiple databases from 1 to N with different features as the inputs.

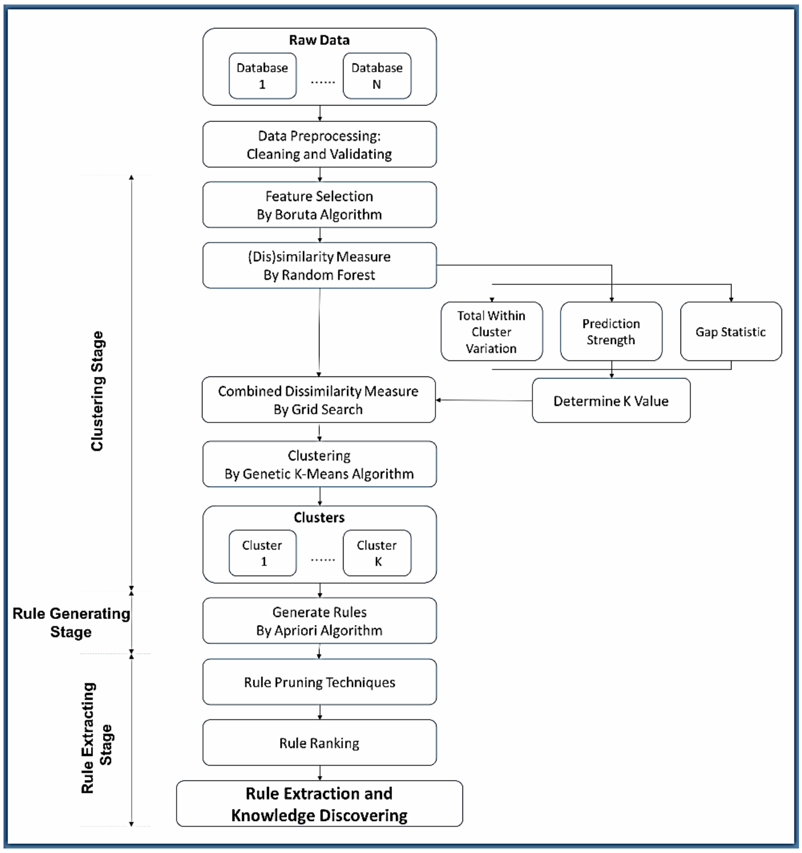
Raw Data
Database 1
……
Database N
Data Preprocessing: Cleaning and Validating
Feature Selection By Boruta Algorithm
(Dis)similarity Measure By Random Forest
Total Within Cluster Variation
Prediction Strength
Gap Statistic
Determine K Value
Combined Dissimilarity Measure By Grid Search
Clustering By Genetic K-Means Algorithm
Clusters
Cluster 1
……
Cluster K
Generate Rules By Apriori Algorithm
Rule Pruning Techniques
Rule Ranking
Rule Extraction and Knowledge Discovering
Clustering Stage
Rule Generating Stage
Rule Extracting Stage

Fig. 1. System architecture

4 Case Study and Results

The data chosen to analyze in this work were collected by the research project *Taiwan Education Panel Survey* (TEPS) [20-22] and *Taiwan Education Panel Survey and Beyond* (TEPS-B) [6]. TEPS is a longitudinal data collected by 4 waves of surveys every two years from 2001 to 2007. It's a questionnaire-based survey targeting students of high school sampled from all the area in Taiwan. Besides the students, the questionnaire was also sent to their parents and teachers, in order to acquire the family and school information. With three sides of surveys, student's status, family and school, TEPS can reflect the holistic environment of a student in the studying process.

In this study, we choose the first wave of TEPS database in our experiments. The first wave of survey was conducted in 2001 targeting students born in 1984 or 1985, and was studying in the second year of high school at that time. More details about the TEPS database can be found in the TEPS user's guide [5].

TEPS-B is a follow-up investigation of TEPS. It's a three-year project conducted from 2009 to the end of 2011. The goal of TEPS-B is to explore the relationship between education and working status by post-tracking students who were sampled in TEPS before. These students were at age 24-25 in 2009 and just entered the labour market for the first or second job. We choose the database from TEPS-B in 2009 since it has more observations sampled than in 2010.

4.1 Data Preprocessing

In this study, we want to discuss how the background of growth will affect one's future achievement. By cascading the first wave of TEPS database with TEPS-B database, one's background in the aspects of family and education can be linked to the working status. Typically, achievement is considered as one's socioeconomic status based on income, education level, and occupation. Since the expansion of higher education in late 1980s, over 85% of people in Taiwan have held bachelor's degree or above. It means that education level was relatively equalized and can't be a representative factor to define socioeconomic status in Taiwan. In addition, due to the restriction of TEPS-B database, the occupations are categorized in an indistinct manner. Hence, income level is chosen to serve as the achievement in this study.

There are 15922 observations in TEPS-B. In the data cleaning and validation step, after eliminating missing, refusal and abnormal value in income, occupation and education, there remain 5622 observations.

As caring more about the factors causing high and low income level, we pick out observations lying in two tails in Fig. 2. Observations are classified as high income group with income level ≥ 11, and classified as low income group with income ≤ 3. There are 135 observations in high income group and 128 observations in low income group which accounts for 2.4% and 2.3% respectively in sample space of 5622 observations. The x-axis is monthly income level ranging from no income to New Taiwan Dollar (NTD) 80000 up with unit increment as NTD 5000. The y-axis is the frequency count of observations in given monthly income level. The selecting principle comes from two reasons. First, the amount of both class labels are expected to be approximately equal since the unbalance of class labels would lower the performance of classification and cause a biased result in mining association rules. Second, the income level is an orderly categorical variable in TEPS-B and distributed in a right skewed manner. To avoid having too many observations as low income group, it’s reasonable to set threshold of income level to be 3.

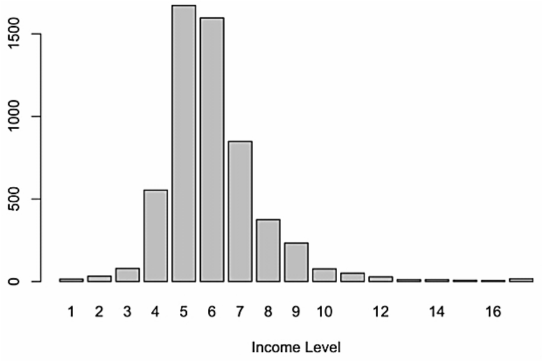

Fig. 2. High and low income level of observations

4.2 Clustering Stage

The clustering stage starts from the feature selection procedure. Setting income level as class label and all other features as independent variables, Boruta algorithm is firstly performed for feature selection with a number of trees equal to 1000 and maximal runs equal to 500. Besides, in order to pick the truly significant attributes from large amount of TEPS feature sets, we take a relatively strict threshold for setting p-value of two-sided test equal to 0.01. There are 9, 4 and 6 features selected correspondingly in student, family and school database by Boruta algorithm. All these 19 features are tested as important features and utilized for further process.

Random Forest algorithm is utilized to derive a supervised similarity measure between each observation. The parameters are set with a number of trees equal to 2000 and a number of variables randomly sampled at each split equal to sqrt(M), where M is the number of features in given database. The OOB (out-of-bag) error rates from classification results are shown in Table 1 by comparing data with all features and only the important features. It shows that the OOB estimate of error rate can be lowered up to 13.7% by conducting classification with selected features. Then, the similarity measures between each observation among the 3 databases are calculated and presented by a symmetric matrix. The dissimilarity is derived from similarity measures as sqrt (1-similarity).

Table 1. The OOB error rates from classification results

OOB Error Rate	Student Database	Family Database	School Database
Without Feature Selection	41.3%	48.3%	43.2%
With Feature Selection	34.6%	34.6%	37.6%

Before applying dissimilarity measure into clustering, a necessary issue is needed to be solving first:

choosing the number of centroids *K*. Total within cluster variation (TWCV), Gap statistic and prediction strength are served as metrics to determine the *K* value. These metrics are derived on the basis of dissimilarity measure by applying RF on TEPS-B database. Fig. 3 shows the TWCV from *K*=2 to 10. With more clusters in the sample space, there are less observations within a single cluster that makes each group become smaller and tighter. Therefore, TWCV continuously decreases with the increasing of *K*. It's recommended to choose *K*=3 by the elbow method.

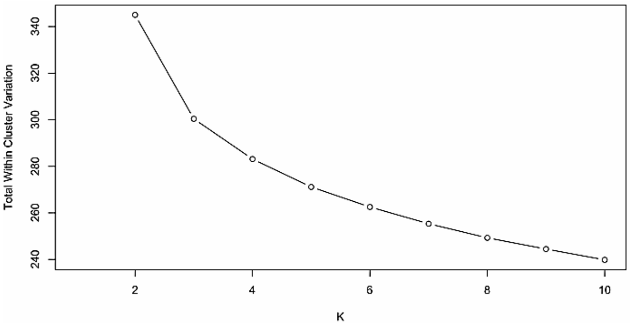

Fig. 3. TWCV from *K*=2 to 10

Fig. 4 depicts Gap statistic (Gapk) with standard error from *K*=1 to 10 by setting the number of observations in reference dataset equal to 150. Higher Gap statistic means there exist relatively more obvious clusters among the given dataset. Normally, *K* is selected with the highest Gap statistic value, i.e., *K*=10. However, it's impractical for *K* to be large since there would be too few observations in a single cluster. From the graph, we pick *K*=3 as it contains the biggest marginal gain.

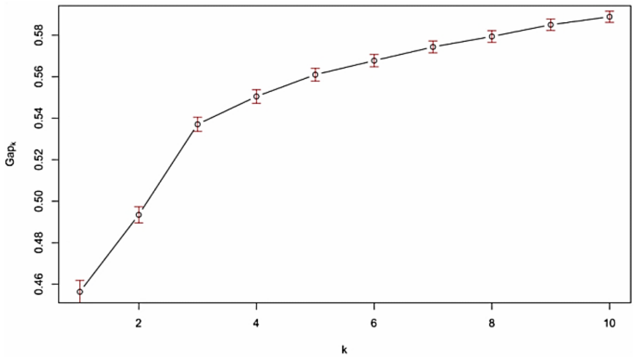

Fig. 4. Gap statistic (Gapk) with standard error

The prediction strength is shown in Fig. 5 with standard error from *K*=1 to 10. High prediction strength implies the correctness of *K* value since the clustering results maintain the same in both training and test dataset. Except for the condition of *K*=1, it seems to be more suitable to select *K*=3 according to the prediction strength value. All of these measurements indicate the *K* value should be chosen to be 3. As a consequence, we adopt the result for the further procedures.

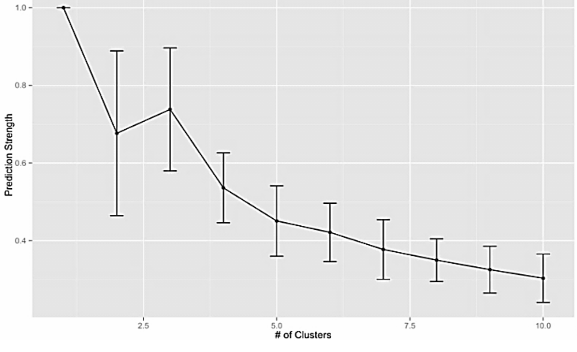

Fig. 5. The prediction strength with standard error from *K*=1 to 10

In order to take consideration of three databases, we assign a weighting value to each dissimilarity matrix and use them to construct a combined matrix. The goal is to find a best set of weighting values that aims to minimize TWCV after clustering. In this parameter optimization problem, grid search algorithm is utilized with performance metric defined as TWCV. Furthermore, the TWCV is derived by Genetic K-means algorithm (GKA) with *K*=3. In GKA, all the observations in dataset are included in initial population and ran for 100 generations with mutation rate = 0.02. After the computation, the best weighting setting is found as (*Student*, *Family*, *School*) = (0.1, 0.6, 0.3). By comparing to TWCV from clustering results ran with each of the database, the combined dissimilarity measure gives the lowest value as shown in Fig. 6. Hence, we can state that instead of looking at each database solely, the clustering results are better when taking all of them into consideration together.

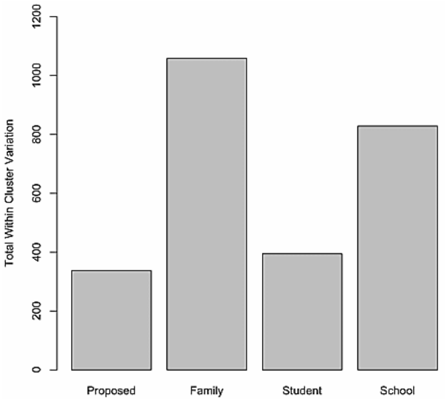

Fig. 6. The combined dissimilarity measure

After performing clustering on combined dissimilarity matrix, the results are visualized by classical multidimensional scaling (MDS). MDS can find a spatial configuration of each object in N-dimensional space with the between-object distances preserved as well as possible. The distances are presented by assigning objects at coordinates in each of the N dimensions. The 3D MDS visualization of clustering results is plotted as Fig. 7. Data points with the same color belong to the same cluster.

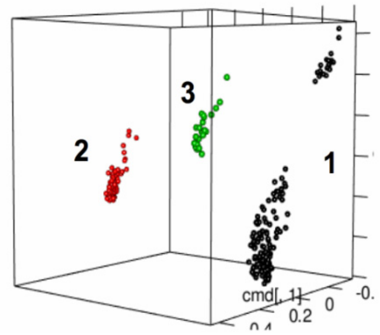

Fig. 7. The 3D MDS visualization of clustering results

From the 3D MDS plot, the three clusters are visually separated in a distinct manner which implies the heterogeneity among clusters. Besides, Table 2 shows the clustering result with income level as class label.

Table 2. The clustering result with income level

No. of Cluster	Income Level = High	Income Level = Low
Cluster 1	56	79
Cluster 2	70	36
Cluster 3	9	13

We want to know if these clusters preserve different characteristics in terms of income level. In order to validate the effectiveness of clustering result, the chi-squared test is utilized to determine the difference of class label observed in each group from the expected. With the degree of freedom being 2, the chi-squared test gives the result of chi-squared measure 15.376 and p-value 0.0004583. The p-value is much lower than 0.05 which indicates the significant difference between observed and expected frequency of class label, i.e., the distribution of class labels with and without clustering. As a consequence, it can be claimed that the clusters do affect the distribution of class labels and can support the existence of heterogeneity among different groups.

4.3 Rule Generation Stage

Apriori algorithm is applied to generate rules among each cluster. As the post-pruning would be performed later, the setting of confidence and support thresholds are less sensitive to the system. Depending on different scenarios, the support threshold would be 0.15 or 0.3 and cardinality ranges from 3 to 4 with confidence threshold = 0.6 or 0.7. The parameters are adjusted based on the number of rules generated such as to constraint the amount of rules within the hecto-scale. We consider the scenarios with three factors: income level, timeline and feature scale.

Income level: The consequents of each rule are confined to class labels as {*Income=High*} or {*Income=Low*}. Since the cluster 1 maintaining more observations with low income labels, the rules indicating {*Income = Low*} as consequent are produced from cluster 1 so to enhance the support values. For the same reason, rules with consequent {*Income = High*} are generated from cluster 2. On the other hand, cluster 3 is left for future analysis in that the observations in the group are too less to be representative.

Timeline: Two scales of timeline are defined as the horizontal and the vertical. For horizontal timeline, we discuss how status-quo affects the current earning. In this scenario, only the features in TEPS-B database are included as the antecedent to generate association rules with income level. For vertical timeline, different time points are connected to figure out the causal relationships. By cascading TEPS and TEPS-B databases, it can show how the past background and experience in high school life would influence one's earning after nine years. The antecedent in rules is composed of factors from student, family and school databases in TEPS.

Feature scale: The distribution of each attributes' value varies greatly that, for some features, one possible value is dominant among all the observations. As a consequence, this kind of features would dominate the results since association rule mining seems support value as one of the main criteria. To jump out of the restriction, the feature scales are formed by two principles, only the important features and all features. By setting only the important features as antecedents, the unimportant and dominant attributes can be pruned to give possibly more meaningful outcomes.

Taking all above perspectives into consideration, the rules are generated under 6 scenarios. In horizontal timeline, two conditions are applied finding relationship between selected TEPS features with consequent being either high income or low income. To note that, we don't look at all features in horizontal timeline since the selected attributes include almost all attributes in TEPS-B database. In vertical timeline, there are four scenarios: two income level versus two feature scale. Each scenario is listed with its abbreviation given in Table 3 and Table 4.

Table 3. High income level versus two feature scales

High Income Level	Horizontal Timeline	Vertical Timeline
All Features	None	HiVerAll
Selected Features	HiHozSel	HiVerSel

Table 4. Low income level versus two feature scales

Low Income Level	Horizontal Timeline	Vertical Timeline
All Features	None	LowVerAll
Selected Features	LowHozSel	LowVerSel

Besides, we employ k-fold cross validation in order to avoid over-fitting. The *k* is selected to be 2 since the larger *k* may result in too less observations in a single set. The setting of association rule generation for support, confidence and cardinality thresholds as well as the average number of rules produced are listed in Table 5.

Table 5. The setting of association rule generation

category	Support	Confidence	Cardinality	#Avg. Rule Produced
LowHozSel	0.15	0.6	4	226
HiHozSel	0.3	0.6	4	276
LowVerSel	0.15	0.6	4	266
HiVerSel	0.3	0.6	4	348
LowVerAll	0.3	0.7	3	6742
HiVerAll	0.3	0.7	3	81810

4.4 Rule Extraction Stage

For extracting relatively meaningful rules, rule pruning techniques are employed with redundant rule pruning as well as chi-squared pruning by setting the threshold X_t=3.84. With the 2-fold cross validation, the average amounts of rules before and after pruned are depicted in Figure 8 and Figure 9. Noteworthy, given the same parameter setting, the scenario of applying all features in vertical timeline with high income label generates extensive amount of rules comparing to other conditions. One of the possible reasons is that most of the features in this condition preserving the concentrated distributions in terms of their values. Our pruning method reduces 80.3% of rules in average, with 97.7% at most in HiVerSel scenario and 32% at least in LowVerAll scenario. Including too many variables sometimes leads to low overlapping rules because that only slight amount of antecedents have intersection between each other.

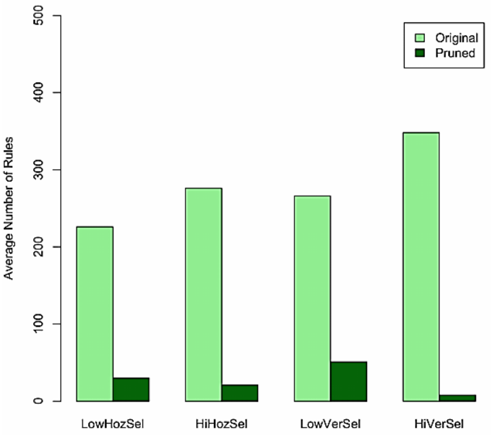

Fig. 8. The average amounts of rules (case 1)

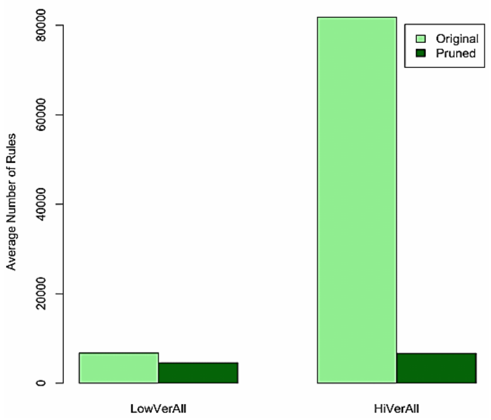

Fig. 9. The average amounts of rules (case 2)

5 Results Analysis and Discussion

The relationship between background factors and earning is presented by two perspectives: Macro and micro viewpoints. In macro viewpoint, the overall impacts from student, family and school database respectively are considered by aggregating the information from all the attributes into three abstract layers. In micro perspective, each attributes from all the databases are involved into discussion and presented by association rules which are what we extracted from the scenarios.

Recall the result from combining different dissimilarity matrixes by Grid search algorithm. The best set of weight values that gives the minimal TWCV value is denoted as (*Family, Student, School*) = (0.1, 0.6, 0.3). A macro view point is provided by considering all the attributes to aggregate into just three ambiguous abstractions. It somehow means that, under the premise to determine current earning by the past background, student-side aspect accounts for 60% of the total components with school-side taken 30% and family-side taken 10%. We interpret the result by follows: As being 11th grade in that time, students were in their adolescent while parents gradually losing the ability to control their children which results in the low influence on students' future. In this age, students' self-worth and behaviours played the most important role on shaping their future achievement with medium impact from school matters.

Next, the micro perspective is provided by stating rules extracted from each scenario. Rules preserving the highest value with either metrics averaged in 2 folds are listed and discussed.

5.1 LowHozSel Scenario

In horizontal timeline, generate rules with consequent as {*Income level = Low*} and selected features as antecedent.

- Rule with the highest chi-squared measure –

  {*Do not have a full-time job now, Past job was not in management level*} → {*Income Level = Low*}. *Support* = 0.22, *Confidence* = 0.88, *Lift* = 1.51, and *Chi-squared measure* = 8.41.
- Rule with the highest confidence and lift –

  {*Do not have a degree of master or doctor, Used to have a job, Past job was not in management level*} → {*Income Level = Low*}. *Support* = 0.19, *Confidence* = 0.9, *Lift* = 1.54, and *Chi-squared measure* = 7.47.
- Rule with the highest support –

  {*Past job was not in management level*} → {*Income Level = Low*}. *Support* = 0.53, *Confidence* = 0.67, *Lift* = 1.14 and *Chi-squared measure* = 8.09.

Rules are trivial in this scenario. Jobless people are more likely to lie into low income group. However, they can be seemed as demonstrations of solving toy problems which denotes the correctness of our model. The education attainment gives impact in this scenario which echoes the finding of related works.

5.2 HiHozSel Scenario

In horizontal timeline, generate rules with consequent as {*Income level = High*} and selected features as antecedent.

- Rule with the highest chi-squared measure –

  {*Have a full-time job now, Hired by a specific firm*} → {*Income Level = High*}. *Support* = 0.52, *Confidence* = 0.84, *Lift* = 1.28 and *Chi-squared measure* = 14.07.
- Rule with the highest confidence and lift –

  {*Hired by a specific firm, Already graduated*} → {*Income Level = High*}. *Support* = 0.46, *Confidence* = 0.86, *Lift* = 1.3, and *Chi-squared measure* = 11.4. *Accuracy* = 0.82.
- Rule with the highest support –

  {*Have a full-time job now*} → {*Income Level = High*}. *Support* = 0.63, *Confidence* = 0.77, *Lift* = 1.18 and *Chi-squared measure* = 13.07.

The results are trivial but with little insight: People graduated from school and owning a job are more likely to get high income. Pursuing education degree and works simultaneously is somehow a more challenging condition to deal with.

5.3 LowVerSel Scenario

In vertical timeline, generate rules with consequent as {*Income level* = *Low*} and selected features as antecedent.

- Rule with the highest chi-squared measure –
  {*Never skip a class, Do not join school teams, Low English proficiency*}→{*Income Level* = *Low*}. *Support* = 0.38, *Confidence* = 0.77, *Lift* = 1.31 and *Chi-squared measure* = 11.42.
- Rule with the highest confidence and lift –
  {*Studied in vocational high school, Low English proficiency*}→{*Income Level* = *Low*}. *Support* = 0.17, *Confidence* = 0.86, *Lift* = 1.47 and *Chi-squared measure* = 5.13.
- Rule with the highest support –
  {*Low English proficiency*}→{*Income Level* = *Low*}. *Support* = 0.56, *Confidence* = 0.61, *Lift* = 1.05 and *Chi-squared measure* = 3.91.

It reflects the inequality of educational resources in different types of schools in our country. Students studied in vocational high school are more likely to achieve worse fulfilment. On the other hand, we can see that English education has a dominant position in Taiwan. Poor English may lead to pessimistic future.

5.4 HiVerSel Scenario

In vertical timeline, generate rules with consequent as {*Income level* = *High*} and selected features as antecedent.

- Rule with the highest chi-squared measure and support –
  {*Joined a high school, Homeroom teacher owns teacher certificate*}→{*Income Level* = *High*}. *Support* = 0.54, *Confidence* = 0.72, *Lift* = 1.09 and *Chi-squared measure* = 4.79.
- Rule with the highest confidence and lift –
  {*The school aimed at entering the higher education or landing a job, Anyone of parents gets hired*}→ {*Income Level* = *High*}.*Support* = 0.39, *Confidence* = 0.76, *Lift* = 1.15 and *Chi-squared measure* = 3.97.

Certain rules are more dominant in this scenario in that they maintain multiple metrics with the highest value. Rules reflect the educational status-quo in Taiwan. Regardless of the policy for promoting vocational high school, students in high school are still holding better chance than others in vocational high school. It indicates the resource inequality in educational facilities. Furthermore, schools targeting to more pragmatic goals, like entering higher education level or getting a job, are expected to assist more on students' future earnings.

5.5 LowVerAll Scenario

In vertical timeline, generate rules with consequent as {*Income level* = *Low*} and all features as antecedent.

- Rule with the highest chi-squared measure and support –
  {*Teachers often help each other (English subject), Never sent to office of student affairs due to inappropriate behaviors*}→{*Income Level* = *Low*}. *Support* = 0.44, *Confidence* = 0.70, *Lift* = 1.2 and *Chi-squared measure* = 6.86.
- Rule with the highest confidence and lift –
  {*Classmates usually chat with teachers, Student never joined a school team*}→{*Income Level* = *Low*}. *Support* = 0.32, *Confidence* = 0.77, *Lift* = 1.31 and *Chi-squared measure* = 6.59.

When taking all features into consideration, rules may be unexplainable due to the counter-instinct combination of antecedents. But there is still something can be dug: to act like a good student and never be sent to office of student affairs due to inappropriate behaviors would not promise you a brighter future. In other words, endurable degree of inappropriate behaviors in school can still be acceptable in a longer time scale.

5.6 HiVerAll Scenario

In vertical timeline, generate rules with consequent as {*Income level* = *High*} and all features as antecedent.

- Rule with the highest chi-squared measure, confidence and lift – {*Still be a teacher if could choose again (English teacher), Father gets hired*}→{*Income Level* = *High*}. *Support* = 0.31, *Confidence* = 0.94, *Lift* = 1.44 and *Chi-squared measure* = 9.52.
- Rule with the highest support – {*Do not group students by their academic performance (Math)*} →{*Income Level* = *High*}. *Support* = 0.63, *Confidence* = 0.73, *Lift* = 1.1 and *Chi-squared measure* = 7.95.

The intention and passion of a teacher can influence a student profoundly in terms of the way the teacher instructs and guides. Teachers with firm aspirations make students move further. Besides, a teacher who would not separate students because of academic performance is more possible to lead them better future.

6 Conclusions

We have established a system including steps as feature selection, clustering and association rule mining. Our system not only picks out the important features but uses them for further analysis such as clustering and association rule mining. The issue of defining distance between objects with categorical attributes is overcome using random forest dissimilarity measure. In addition, dissimilarity measures from different databases are combined with the aid of the optimization method. By choosing the number of clusters in a systematic manner, the clustering results are proven to be statistically significant. Rule pruning and ranking approaches also make the meaningful rules in each cluster being extracted successfully

The finding of our studies can be examined in macro and micro perspectives. We found that the determinants of one's earning would vary depending on macro or micro perspectives which are rarely mentioned in other works. In a macro viewpoint, the factors influenced future earning is aggregated into three abstract layers representing self-worth, family background and school life. We found that, in one's 11th grade, the student herself/himself affects her/his future earning most, while his/her family background being the least one. On the other hand, in a micro perspective, rules supported by significant metrics are inspected in 6 scenarios in terms of income level, timeline and feature scales. While some of rules are supportive to the previous work as well as some are trivial, there are still unexpected and surprising insights found by our approach. These findings on one hand reflect the status-quo of our society, and on the other hand can guide young students to react for aiming a high-earning future.

We would like to note that a limitation of our work is that it can't have more data sets. The future works are proposed in some possible ways. By setting different attributes as consequents, the relationship of various factors can be investigated in-depth. For cascading more databases surveyed in different years, the time series can be included and examine the changing within diverse time scale. Last, other machine learning techniques can be implemented in our system to enhance the reliability and robustness.

References

[1] R. Rumberger, Education and the reproduction of social inequality in the United States: an empirical investigation, Economics of Education Review 29(2)(2010) 246-254.

[2] M. Doepke, M. Tertilt, Families in Macroeconomics, Handbook of Macroeconomics, Elsevier, Amsterdam, Netherlands, 2016.

[3] S. Bowles, H. Gintis, M. Osborne, The determinants of earnings: a behavioral approach, Journal of Economic Literature 39(4)(2001) 1137-1176.

[4] A.J. Izenman, Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning, Springer, New York, 2013.

[5] L.-Y. Chang, Taiwan education panel survey: users' guide and the first wave, Center for Survey Research, Academia Sinica, 2003.

[6] P.-Y. Kuan, Taiwan education panel survey and beyond/2009: telephone follow-up survey of panel-1 SH, Survey Research Data Archive, Center for Survey Research, Research Center for Humanities and Social Sciences, Academia Sinica, 2014.

[7] W.-Y. Loh, Classification and regression trees, WIREs: Data Mining and Knowledge Discovery 1(1)(2011) 14-23.

[8] M.B. Kursa, A. Ankowski, W.R. Rudnicki, Boruta: a system for feature selection, Fundamenta Informaticae 101(4)(2010) 271-285.

[9] M. Okabe, S. Yamada, Clustering using boosted constrained k-means algorithm, Frontiers in Robotics and AI 5(2018), Article18. Doi:10.3389/frobt.2018.00018.

[10] D. Zeebaree, H. Haron, A. Abdulazeez, S. Zeebaree, Combination of K-means clustering with genetic algorithm: a review, International Journal of Applied Engineering Research 12(24)(2017) 14238-14245.

[11] P. Bholowalia, A. Kumar, EBK-Means: A Clustering Technique Based on Elbow method and K-Means in WSN, International Journal of Computer Applications 105(9)(2014) 17-24.

[12] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society 63(2)(2001) 411-423.

[13] R. Tibshirani, G. Walther, Cluster validation by prediction strength, Journal of Computational and Graphical Statistics 14(3)(2005) 511-528.

[14] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proc. 1994 International Conference on Very Large Data Bases, 1994.

[15] R. Topor, H. Shen, Construct robust rule sets for classification, in: Proc. 2001 International Conference on Knowledge Discovery and Data Mining, 2001.

[16] F. Thabtah, A review of associative classification mining, The Knowledge Engineering Review 22(1)(2007) 37-65.

[17] R.L. Ott, M. Longnecker, An Introduction to Statistical Methods and Data Analysis, Cengage Learning, Brooks/Cole, 2016.

[18] Y. Ye, T. Li, Q. Jiang, Y. Wang, CIMDS: adapting post processing techniques of associative classification for malware detection, IEEE Transactions on Systems, Man, and Cybernetics 40(3)(2010) 298-307.

[19] M. Antonie, O. Zäıane, A. Coman, Associative Classifiers for Medical Images, Mining Multimedia and Complex Data, Springer-Verlag, Berlin, Germany, 2003.

[20] L.-Y. Chang, Taiwan education panel survey: the first wave (Student Data) (C00124_A), Survey Research Data Archive, Taiwan Academia Sinica, 2003.

[21] L.-Y. Chang, Taiwan education panel survey: the first wave (Teacher Data) (C00124_C), Survey Research Data Archive, Taiwan Academia Sinica, 2003.

[22] L.-Y. Chang, Taiwan education panel survey: the first wave (Parent Data) (C00124_G), Survey Research Data Archive, Taiwan Academia Sinica, 2003.