The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2398-7294.htm

Knowledge discovery in sociological databases

An application on general society survey dataset

Zhiwen Pan
Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
Jiangtian Li
High School Affiliated to Renmin University of China, Beijing, China
Yiqiang Chen
Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
Jesus Pacheco
Universidad de Sonora, Hermosillo, Mexico, and
Lianjun Dai and Jun Zhang
Information Centre of China Disabled Persons' Federation, Beijing, China

Received 12 September 2019
Revised 11 October 2019
Accepted 12 October 2019

Abstract

Purpose – The General Society Survey(GSS) is a kind of government-funded survey which aims at examining the Socio-economic status, quality of life, and structure of contemporary society. GSS data set is regarded as one of the authoritative source for the government and organization practitioners to make data-driven policies. The previous analytic approaches for GSS data set are designed by combining expert knowledges and simple statistics. By utilizing the emerging data mining algorithms, we proposed a comprehensive data management and data mining approach for GSS data sets.

Design/methodology/approach – The approach are designed to be operated in a two-phase manner: a data management phase which can improve the quality of GSS data by performing attribute pre-processing and filter-based attribute selection; a data mining phase which can extract hidden knowledge from the data set by performing data mining analysis including prediction analysis, classification analysis, association analysis and clustering analysis.

Findings – According to experimental evaluation results, the paper have the following findings: Performing attribute selection on GSS data set can increase the performance of both classification analysis and clustering analysis; all the data mining analysis can effectively extract hidden knowledge from the GSS data set; the knowledge generated by different data mining analysis can somehow cross-validate each other.

This work is supported by the National Key R&D Program of China (2018YFC2002603) and CCF-NSFOCUS KunPeng Research Fund (2018013).

International Journal of Crowd Science
Vol. 3 No. 3, 2019
pp. 315-332
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-09-2019-0023

Originality/value – By leveraging the power of data mining techniques, the proposed approach can explore knowledge in a fine-grained manner with minimum human interference. Experiments on Chinese General Social Survey data set are conducted at the end to evaluate the performance of our approach.

Keywords Data management, Data mining, Crowdsourced big data and analytics, Knowledge discovery

Paper type Research paper

1. Introduction

The General Society Survey (GSS) aims at examining the socio-economic status and structure of contemporary society, as well as the living condition and public opinions of the individual subjects (Davis and Smith, 1991). The surveys are conducted annually by non-profit professional institutes (e.g. research center in the university). The process of conducting GSS include designing sociological survey forms with a systematic set of questions; distributing the forms to social subgroups or social individuals; authenticating and integrating the survey feedbacks as tabular data set; release the data set to public. As the GSS is usually conducted with a big amount of profession effect and is funded by the government, the GSS data set is regarded as one of the authoritative source for making data-driven policies. That is, by analyzing the GSS data set, practitioners in government and other social organizations can have a better understanding regarding the status and demands of the social groups to be served, so that rational policies can be made accordingly.

Currently, nearly all of the researches on GSS data set are conducted by following the classic sociological research methods where expert knowledges and statistics are combined. In general, these methods determine the variable inputs of each analysis based on expert knowledges and casual inferences, and then conduct analysis by using statistical algorithms (Hu and Leamaster, 2015; Tan, 2014; Wu *et al.*, 2014; Johnston, 2017). These methods have some inherent drawbacks to be addressed:

- The researches based on classic methods are time-consuming and require many expert efforts.
- As the GSS data set consist of a huge amount of data attributes which are correlated and interactional with each other, traditional statistics algorithms are not intelligent enough to fully explore these correlations and interactions between the data attributes. Consequently, the research results been generated may be simple and limited.
- As GSS data set are secondary data set which means the survey forms are designed by the researcher themselves, the data attributes within the data sets may not be adequate for researcher to create models based on their knowledges and inferences (Johnston, 2017).
- There is no comprehensive analysis approach which can perform GSS data analysis in variety kinds of perspectives.

By leveraging the power of data management and data mining techniques to the analytics of GSS data set, we propose a Comprehensive Big Data Management and Mining (CBDMM) approach which can provide the professions a unified and effective way to analyze the GSS data sets. The CBDMM approach works in a two-phase manner: data management phase and data mining phase. During the data management phase, GSS data pre-processing method and filter-based attributes selection method are designed to improve data quality and filter out redundant attributes for the oncoming data mining analyses. During the data mining phase, a set of state-of-art data mining algorithms are implemented to perform holistic analyses including classification analysis, association analysis, predictive analysis and clustering analysis. The results generated by these data mining analyses can be interpreted and visualized, and then be stored as knowledge.

The proposed approach can address the aforementioned drawbacks as follows: the approach rely on data mining algorithms to manage and analyze data, it requires less professional effort and is time-efficient; due to the advancement of the attribute selection method and data mining algorithms, our approach can generate more exhaustive and accurate results than traditional approach; the approach can provide comprehensive analysis with algorithms in different analytic categories, hence it can be used to perform variety kinds of analytic tasks.

The rest of this paper is organized as follows: In Section 2, the data management phase of our proposed CBDMM approach is introduced. In Section 3, the data mining phase of our proposed CBDMM approach is introduced. In Section 4, the data management and mining results are presented to prove the effectiveness of our approach. Finally, our works are summarized in Section 5.

2. Comprehensive big data management and mining approach

As a kind of authoritative survey data set that is designed with a lot of professional effort, the GSS data sets obtain the following characteristics:

- High data dimension: A GSS data set contains hundreds of data attributes which are corresponded to hundreds of survey questions.
- High data quality: The anomaly data and low-quality data samples (e.g. missing values) are filtered during GSS data gathering.
- Heterogeneous data: the GSS data set contains continuous-valued attributes, discrete-valued attributes and enumerated-valued attributes.
- Abundant information: The data attributes within GSS data set describe the status, conditions and opinions of social individuals in variety of perspectives.

According to the characteristics of GSS data set, our CBDMM framework is designed to operate in two-phase manner, the detail operations of each phase are shown in Figure 1. The attribute pre-processing in data management phase aims at dealing with heterogeneity of GSS data attributes so that data mining analysis can be performed with valid attribute values. The attribute selection in data management phase aims at dealing with high data dimension of GSS data set. Specifically, by applying a set of feature selection algorithms and fusing the selection result with aggregation algorithms, sub-data sets where attributes are high correlated are generated for different data mining tasks. The prediction analysis, classification analysis, association analysis, and cluster analysis in data mining phase aim at recognizing the correlations between the data attributes so that hidden information within the data set can be extracted. These analyses perform correlation recognition in different perspectives. In this way, a rich set of hidden information can extracted for different research topics, and the information extracted from different analysis can be mutually conformed among each other. The result interpretation in data mining phase aims at interpreting and visualizing the hidden information into knowledges that can be easily understood by decision makers. The interpretation and visualization methods vary for different kinds of analysis (even vary for different algorithms). The detail of CBDMM approach will be introduced in the following sections.

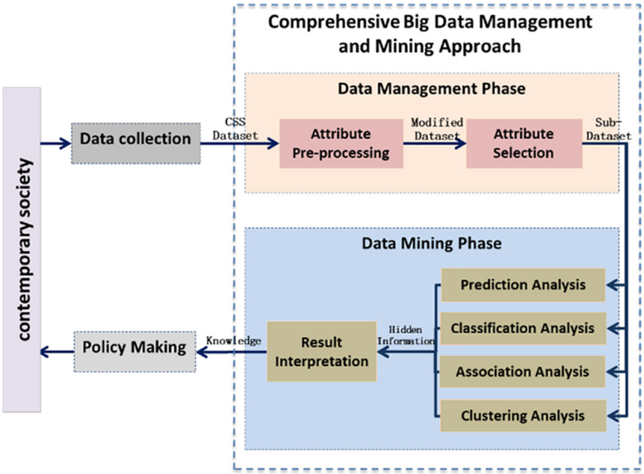


Figure 1.
The architecture of comprehensive big data management and mining approach

3. Data management phase

3.1 Attribute pre-processing

The GSS data set contains three types of data attributes: continuous-valued attributes, discrete-valued attributes and enumerated-valued attributes. The value of continuous-valued attributes is integer or float numbers. Such attributes are usually used to indicate the subjects' age, income and expense. Although continuous attribute values may contain more details than other types of attributes values (e.g. enumerated attributes values), sometimes these details may be unnecessary. Moreover, continuous-valued attributes are not suitable for doing some kinds of analyses such as association analysis. Therefore, attribute pre-processing for continuous attribute values is to convert the continuous values into discrete values by separating the values into a certain intervals. The separations of continuous values are performed based on a certain well-recognized scales. For instance, the "personal income" attribute values can be categorized into high, medium, medium low and low incomes, based on the national income tax rates (Riccardi and Lorenzo, 2013). The "age" Attribute values can be categorized into children (0-14 years), Youth (15-24 years), Adult (25-64 years), Seniors (65 years and over) (Statistics Canada, 2017). Other age categories are list as follows (Du and Yang, 2010; Australian Bureau of Statistics, 2014).

The enumerated valued attribute indicates subjects' selection of a set of enumerated options for a question. Unlike discrete attribute values, there is no rand-size relationship between enumerated attribute values. Hence, enumerated valued attributes without preprocessing are only suitable for doing association analysis and (rule-based) classification analysis. The attribute pre-processing for enumerated valued attributes is to convert these attributes to the other two types of attributes. In this paper, we use one-hot coding to represent enumerated valued attributes, so that these attributes can be converted to discrete valued attributes. Specifically, we represent a question with n choices as a one-hot attribute array: $X_m = \{X_{m,1}, X_{m,2} \cdots X_{m,n}\}$, where each one-hot attribute has one to one correspondence to each choice.

3.2 Attribute selection

Researches on GSS data set are conducted through determining the target classes based on research topics, and then performing data analyses based these target classes. For instance, if the research topic is to explore the characteristic of the high-income population, the attribute value "high income" may be selected as the target class, and then a set of data analyses will be performed accordingly. However, since the GSS data set is a high-dimensional and general data set, there are always some redundant data attributes which are less relevant to the target classes. These redundant data attributes may not only lower the accuracy of data analyses but also increase the time and resource consumption of data analyses. To automatically filter out these redundant features, we proposed a filter-based attribute selection approach (see Figure 2). The approach is operating in a two-step manner. In the first step, a set of attribute selection criterion are used to generate rank lists which can rank the relevance of the attributes. In the second step, a set of rank aggregation algorithms are used to fuse the ranks as a consensus rank. In this way, a sub-data set which only includes the high-relevance attributes can be generated for further data mining analysis.

The criterions we used to rank the relevance of attributes are as follows:

3.2.1 Spectral score (SPEC). This criterion use the Radial Basis Function (RBF) to quantify the similarity between data samples. That is, given two samples x_i and x_j, their similarity between each other can be calculated by the following equation.

$$S_{ij} = e^{\frac{-\|x_i - x_j\|^2}{2\alpha^2}} \tag{1}$$

By calculating the similarity of all pairs of data samples, a corresponding similarity graph with n vertices corresponded to n data samples is established. Within the graph, the links between vertices are weighted similarity between them. The relevance of data attributes can then be ranked based on the spectrum of the graph (Zhao and Liu, 2007).

3.2.2 Information score (IS). This criterion is proposed based on the information theory where the intra-group randomness is low when the similarity of data samples within the group is high. Such similarity can be quantified through entropy. By using the RBF to calculate the similarity S_{ij} between data sample, the entropy value of data samples can be calculated as follows (Mitra *et al.*, 2002):

$$\mathrm{E} = -\sum_{i=0}^{n}\sum_{j=0}^{n} S_{ij} log_2 S_{ij} + (1 - S_{ij}) log_2 (1 - S_{ij}) \tag{2}$$

The relevance of a data attribute can be estimated by observing the reduction of entropy resulted from eliminating the data attribute from the data set.

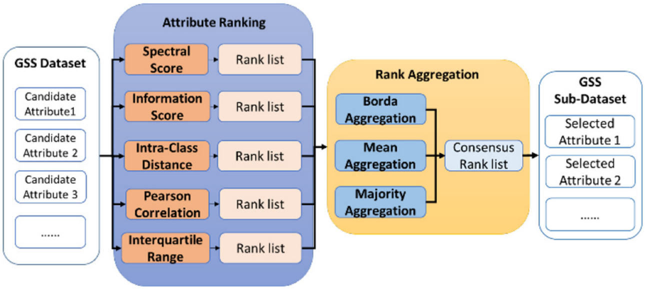


Figure 2.
Flow diagram of attribute selection approach

3.2.3 Pearson correlation. Pearson correlation is a statistic which can reflect the degree of linear correlation between two variables. The correlation between two attributes X and Y can be calculated based on the following equation.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\ \sigma_Y} = \frac{E[(X - \ \mu_X)(Y - \ \mu_Y)]}{2a\sigma_X\ \sigma_Y} \tag{3}$$

where cov (X, Y) is the covariance of X and Y, σ_x and σ_Y are the standard deviation of X and Y. μ_X and μ_Y are the mean value of X and Y, The value of $\rho_{X,Y}$ is range between −1 and 1, and its absolute value is the quantification of inter-attribute correlation. Therefore, the correlation of data attributes x_i can be calculated by adding up its correlation with other attributes:

$$corr(f_i) = \sum_{j=1}^{m} \left| \rho_{x_i,x_j} \right| \tag{4}$$

where m is the number of data attributes within the data set.

3.2.4 Intra-class distance. By adopting Standard Euclidean Distance measurement, this criterion measure the distance from all data samples to the centroid of the class. Providing a class (e.g. high-income population class), the ICD can be calculated as (Kruidenier *et al.*, 2009):

$$IE = \frac{1}{n}\sum_{i=0}^{n} d\left(x_i - x'\right) \tag{5}$$

where x′ is the centroid of the class and x_i is data sample within the class. The relevance a data attribute can be estimated by observing the reduction of ICD resulted from eliminating the data attribute from the data set.

3.2.5 Interquartile range (TR). This criterion is designed based on the principle that the attributes whose values within a class are tightly distributed are more likely to be closely correlated to that class. To quantify such distribution density, the interquartile range scale is adopted. That is, providing a class, the mean interquartile of attribute values corresponding to that class is calculated, and the attribute with lower mean interquartile will be considered as an attribute with higher correlation.

Based on these attribute selection criterion, attribute ranks which indicating the priority of each attributes are generated. The next step is to conduct an aggregation of these ranks, so that a consensus attribute rank can be obtained. In this paper, we propose a set of rank aggregation methods to compare their aggregation performance. During the GSS data analysis, rank aggregation method which can generate sub-data set with higher classification analysis accuracy should be used.

3.2.6 Mean aggregation. Mean aggregation method is to calculate the average rank of each attribute in all rank lists, and take this value as the basis to determine the consensus rank (Dittman *et al.*, 2013). That is, the attributes with higher average rank will be ranked higher in the consensus rank.

3.2.7 Borda aggregation. Borda aggregation method is to averaging the position of features in each rank list. Specifically, the position of an attribute within a rank is described as the number of attributes whose ranks are lower than the target attribute (Dwork *et al.*, 2001). Accordingly, the Borda score of an attribute is the sum of the attribute's position for the all the rank lists. The attributes with higher Borda score will be ranked higher in the consensus rank.

3.2.8 Majority aggregation. Majority aggregation method is proposed based on the majority voting rule. In this paper we utilized Boyer–Moore majority vote algorithm to perform such aggregation. That is, taking the position of an attribute within all the rank list as a set of vote, the vote with highest number will be selected as the position of the attribute in the consensus rank (Dwork *et al.*, 2001).

4. Data analytics phase

4.1 Association analysis

The Association Analysis aims at exploring the association and cause-and-effect relationship between different attribute values. Such exploration is made by finding the attribute values which always appear together. In this paper, we perform association analysis by using a machine learning algorithm named FP-growth (Borgelt, 2005).

The FP-growth algorithm is implemented to be operated in three steps. The first step is to delete all the infrequent attribute values from the data sample and then reorder the frequent attribute values within all the data samples based on their frequency of occurrence. The frequencies of attribute values can be quantified based on their support value. That is, given a data set matrix $T_{n\times m}$, the support level of each attribute value x_i within the data set is calculated as follows:

$$\mathrm{supp}(x_1) = \frac{\left|\{t \in T; x \in t\}\right|}{|T|} \tag{6}$$

where t is a row within matrix $X_{n\times m}$ and $|\{t \in T; x \in t\}|$ is the number of rows which contain the attribute value x. By setting a pre-defined support value threshold, all the attribute values whose support values are below the threshold will be regarded as infrequent, and will be remove from their corresponding data samples. The frequent data attributes values within each data sample will then be reordered according to their support values. In the second step, the Frequent Pattern tree (FP-tree) is built to describe the distribution of data samples in a structured manner. A FP-tree is illustrated as an instance in Figure 3. Being initialized with empty set $\varnothing$, the tree graph stores all the attribute values which are located in the first position of data samples in the first layer of FP-tree. Attributes in the second position will then be stored in the second layer. Such operation is repeated until all the attributes are stored into the tree. The third step is to extract associate rules from the FP-tree. For each node within the tree, the corresponding paths named as Conditional Pattern Base (CPB) can be extracted. These CPB will be selected as the associated rules

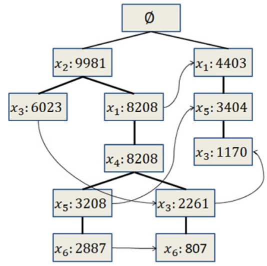


Figure 3.
A sample FP-tree generated by association analysis

based on a certain threshold. Other than support value, another threshold being widely used is confidence level which can be calculated as follows:

$$conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X) \tag{7}$$

where X and Y are frequent attribute value sets (e.g. an association rule $\mathrm{X} \Rightarrow \mathrm{Y}$ can be $\{\mathrm{x}_1, \mathrm{x}_5\} \Rightarrow \{\mathrm{x}_3\}$). In this way, the CPB with confidence level higher than the pre-defined threshold will be selected as the association rules.

4.2 Classification analysis

The classification analysis for data mining aims at exploring the pattern difference between data sets that belong to different classes. Although classification algorithms are developed for decades and are widely used in many areas, most of classification model been generated (such as neural networks) cannot be interpreted. Moreover, for these interpretable models (such as decision trees), the classification accuracy cannot be guaranteed and the extraction of patterns (the knowledges we need) from the model have to be conducted manually which requires many professional efforts. Therefore, in this paper, we implement a powerful interpretable classification algorithm named Random Forest to achieve higher classification accuracy, and implement an algorithm named DPclass to automatically extract rule-based patterns from the Random Forest model.

By choosing target classes and performing data training with Random Forest algorithm, classification model which can distinguish data samples in different classes are generated. The classification model consists of an ensemble of Decision Tree classifiers. Each classifier can generate a classification decision for the current data sample. A final decision can be made based on a vote mechanism so that a decision supported by a majority of the classifiers can be selected (Gao *et al.*, 2017). Compared with other classification algorithms, it has been well proved by researchers that random forest algorithm can achieve better performance on high-dimensional data sets.

After the data training, we use DPclass algorithm to extract discriminative patterns from the Random Forest model (Gao *et al.*, 2017). The major principle of DPclass is searching the local optimal solution iteratively and fusing these solutions to approximate the global optimal pattern. Instead of considering different patterns independently, DPclass can extract patterns based on the effects of the pattern combinations. Therefore, compared with other methods, the pattern extracted by DPclass are more interpretable and less overlapping. The patterns extracted by DPclass refers to a conjunction clause of conditions on feature dimensions which can be defined as follows:

$$P_n = (x_{i,j_1} < v_1) \cap (x_{i,j_2} \geq v_2) \ldots \cap (x_{i,j_m} \geq v_m) \tag{8}$$

where x_i is the i-th data attributes, j is the specific dimension, v indicates the threshold value and m is number of feature within this pattern. After collecting the instances from non-leaf nodes of the decision trees, the patterns extract from every prefix path from the root of a tree to its non-leaf node. Eventually, the patterns that describe the difference between classes can be obtained from random forest model, and these patterns will be represented as a set of rules.

When performing classification analysis on large-scale data set such as CGSS data set, redundant patterns might be extracted. To ensure the qualities of the output patterns, we apply the forward pattern selection to find the rules with higher quality (e.g. top-k patterns). We define the forward selection function as:

$$f_s\colon \{0,1\}^{|m|} \rightarrow \{0,1\}^{|k|} \tag{9}$$

where m is the dimension of the discriminative pattern, and k is the number of the patterns we want to get. To find top-k rules, the forward selection function needs to run iteratively for k times. At each iteration, each discriminative pattern p that not in the discriminative pattern set P_k will be traversed. At the end of each iteration, find the p that maximizes the accuracy of the prediction and add it into the old P_k. In this way, we get the final set P_k which contains top-k discrimination patterns we need.

4.3 Clustering analysis

The clustering algorithms aims at grouping the data samples into cluster a way that the data distribution pattern of data samples within the same cluster are more similar compared with samples in other clusters. The clustering analysis can be used to classify the social population into groups in an unsupervised manner and explore the inherent similarity of these social groups.

Most of clustering algorithms cannot extract interpretable patterns from data set, as the boundaries of clusters generated by these clustering algorithms are irregular and cannot be described with interpretable patterns. In this paper, by using the discriminative rectangle mixture model algorithm (DReaM) (Tibshirani, 1996), clusters with rectangle-shaped boundaries can be generated. In this way, the cluster boundary can be described by interpretable patterns which are formed as decision rules. A vector $t_{\mathrm{kd}} = \left[t_{\mathrm{kd}}^{-}, t_{\mathrm{kd}}^{+}\right]^{\mathrm{T}}$ was defined to represent the decision boundaries, where k is the num of the cluster and d means dimension. The algorithm is designed by assuming that t_{kd} is a sample from a multivariate Gaussian distribution, $t_{\mathrm{kd}} \sim \mathrm{N}(\tilde{\mu}_{\mathrm{kd}}, \tilde{\sigma}_{\mathrm{kd}})$, where $\tilde{\mu}_{\mathrm{kd}}$ is the mean and $\tilde{\sigma}_{\mathrm{kd}}$ is the covariance. If sample x_n satisfies $t_{\mathrm{kd}}^{-} < \mathrm{x}_n < t_{\mathrm{kd}}^{+}$, x_n is inside the decision rectangle and belongs to cluster k. Hence, we can get a prior distribution probability $p(t_{\mathrm{kd}})$ to each decision boundary t_{kd} based on the equation below.

$$\mathrm{p}(t_{kd}) \propto exp\left(-\frac{1}{2}\ \alpha_t\left(t_{kd}^{-} -\ \mu_{t_{kd}^{-}}\right)^2 - \frac{1}{2}\ \alpha_t\left(t_{kd}^{+} -\ \mu_{t_{kd}^{+}}\right)^2 - \frac{1}{2}\beta_t\left(t_{kd}^{+} - t_{kd}^{-}\right)^2\right) \tag{10}$$

In this way, DReaM can approximates the boundary via variational inference, and at the end generated rectangular decision rules which can explicitly describe the boundary of the cluster. Through describing the cluster boundaries, the rectangular decision rules reflect the difference between different clusters. These differences can be interpreted as the knowledges to assist policy-making.

4.4 Prediction analysis

The Prediction Analysis aims at evaluating the influence of some attributes to the variation of a certain numerical attributes within disability data sets. The evaluation is performed by predicting the variation trend of the target numerical attribute and determining the weighted factors of other attributes for the variation trend. Predictive learning is a common application in data mining. The purpose is to predict the unknown values using the known attributes. The linear regression algorithm assumes that the relationship between the predictive value y and the attributes $\mathrm{x} = (\mathrm{x}_1,\mathrm{x}_2,\cdots,\mathrm{x}_n)$ is linear. It takes the form Eq(8).

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n + \ \varepsilon \tag{11}$$

where $\boldsymbol{\epsilon}$ is the error term that captures all other factors which influence variable y other than the attributes **x.** However, the linear regression model doesn't consider interactions between features. The RuleFit algorithm was proposed by Friedman and Popescu in 2005 (Friedman and Popescu, 2008). It could solve this problem by learning sparse linear models to capture interactions between the original features in the form of decision rules.

The RuleFit algorithm could generate some rules such as IF $x_1 \in \{1,2,3\}$ *AND* $x_2 < 4$ *THEN 1 ELSE 0*. A tree ensemble could be used to convert into multiple rules. Any path to a node in a tree is regarded as a decision rule. The tree ensemble model can be described as:

$$\mathrm{F}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m f_m(\mathbf{x}) \tag{12}$$

where M is the number of trees and each ensemble member $\mathrm{f_m}(\mathbf{x})$ is the prediction function of the *mth* tree. $F(\boldsymbol{x})$ is the linear combination of prediction function with the weights $\{\mathrm{a_m}\}_0^{\mathrm{M}}$.

The RuleFit creates the rules $r_m(\boldsymbol{x})$ instead of prediction function $f_m(\boldsymbol{x})$. It generates a new set of features from original features. We assume that $\mathrm{S_j}$ is the set of all possible values for input x_j, and S_{jm} is a subset of S_j. Each rule is defined as

$$r_m(\boldsymbol{x}) = \prod_{j=1}^{n} I(x_j \in s_{jm}) \tag{13}$$

where n represents the number of new features used in the *mth* tree, I(·)is an indicator function that is 1 when feature $\mathrm{x_j}$ is in the subset of the *jth* original features and 0 otherwise. These features are binary r_{m} (x) ∈ {0,1}. As to GSS data set, the features include numerical features and categorical features. For numerical features, S_{jm} is an interval like $20 \leq x_{age} \leq 30$. For categorical features, $\mathrm{S_{jm}}$ is the subset of some specific categories like $x_{investment} \in \{stock, fund\}$.

The total number of rules created from an ensemble of M trees with t_{m} terminal nodes each is

$$\mathrm{N} = \sum_{m-1}^{M} 2(t_m - 1) \tag{14}$$

The predictive model is changed from equation (9):

$$\mathrm{F}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m r_m(\boldsymbol{x}) \tag{15}$$

with the loss function:

$$\{a_m\}_0^M = argmin_{\{a_m\}_0^M} \sum_{i=1}^{N} L\left(y_i, a_0 + \sum_{m=1}^{M} a_m r_m(\boldsymbol{x_i})\right) + \ \lambda \cdot \sum_{m=1}^{M} \left|a_m\right| \tag{16}$$

where N represents the number of training data. y is the truth value and x is the joint values of input variables. The first term in Eq(13) measures the prediction risk on the training sample using a variety of loss functions $\mathrm{L}(\mathrm{y}, \hat{y})$. And the second (regularization) term penalizes the sum absolute values of model parameters. λ is regularization parameter that determines the relative importance of keeping the model simple relative to reducing training error. When λ is zero then the regularization term becomes zero. We are back to the original loss function.

To improve the robustness against input variable outliers, we winsorize the original features. Every rule and every original feature becomes a feature in the linear model because trees couldn't represent simple linear relationships between y and x.

$$l_j(x_j) = min\left(\delta_j^+, max(\ \delta_j^-, x_j)\right) \tag{17}$$

where δ_{j}^+ and δ_{j}^- are the δ quantiles of the data distribution $\{x_{ij}\}_{i=1}^N$ for each feature $\mathrm{x_j}$. Generally, small values (δ = 0.025) are sufficient. $\mathrm{x_j}$ that is in the 2.5 per cent lowest or 2.5 per cent highest values will be set to the quantiles at 2.5 or 97.5 per cent. respectively. We add these additions to train the predictive model with Lasso, with the following structure

$$\mathrm{F}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m r_m(\boldsymbol{x}) + \sum_{j=1}^{p} b_j l_j(x_j) \tag{18}$$

Since RuleFit uses the Lasso, the loss function gets the additional constraint

$$\left(\{a_m\}_0^M, \{b_j\}_1^p\right) = argmin \sum_{i=1}^{N} L\left(y_i, a_0 \sum_{m=1}^{M} a_m r_m(\boldsymbol{x}_\mathbf{i}) + \sum_{j=1}^{p} b_j l_j(x_{ij})\right) + \ \lambda \cdot \left(\sum_{m=1}^{M} \left|a_m\right| + \sum_{j=1}^{p} \left|b_j\right|\right) \tag{19}$$

In addition, the linear term in Eq(18) could be normalized to guarantee the same prior importance as a decision rule.

$$l_j^*(x_j) = 0.4 \cdot l_j(x_j)/std\left(l_j(x_j)\right) \tag{20}$$

The term $\mathrm{l_j^*}(\mathrm{x_j})$ is used in equations (17) and (18) Here $\mathrm{std}\left(\mathrm{l_j}(\mathrm{x_j})\right.$ is the standard deviation of $\mathrm{l_j}(\mathrm{x_j})$ and 0.4 is the average standard deviation of rules with a uniform support distribution of $\mathrm{S_k} \sim \bigcup(0,1)$.

The interpretation of the RuleFit algorithm is similar to linear model. The difference is that model generates new binary features derived from decision rules. The RuleFit outputs all predictive functions (i.e. rules and linear functions) included in the ensemble, with their respective coefficients. The coefficients represent the increase in the predicted value for a unit increase in the predictive function. As seen in equation (15), the predicted output F(**x**) changes by $\mathrm{b_j}$ if feature $\mathrm{x_j}$ changes by a unit and other features remain unchanged. Similarly, in the first linear term, if all conditions of a decision rule $r_{\mathrm{m}}(\mathbf{x})$ apply, the predicted output changes by the learned weight $\mathrm{a_m}$.

5. Experimental evaluation

5.1 Experimental data set

To prove the effectiveness of our proposed approach, we implemented our approach to analyze China General Social Survey (CGSS) data set, which is the earliest national representative continuous survey project run by academic institution in China (NSRC, 2019). CGSS is aimed to systematically monitor the relationship between social structure and quality of life in Chinese society (NSRC, 2019). We choose to analyze the open-source 2015 CGSS data set which contains 10968 data samples collected from 10968 individuals. Within the survey, there is a question: "In general, Do you think you live in a happy life?" The Corresponding answers are "Very unhappy", "Relatively unhappy", "Not sure", "Relatively happy", "Very happy". The major topic of this experiment is to explore the relationship between society individuals' life happiness and their living conditions and attitudes. To simplify the classification task, we classify the degree of life happiness into two categories: happiness or unhappiness. Specifically, the two answers "very unhappy" and "relatively unhappy" fall into the category of unhappiness. The two answers "very happy" and "relatively happy" fall into the category of happiness. The answer "Not sure" are discarded during our experiments. With the help of the domain experts, we selected 45 attributes from the CGSS data set which might be relevant to society individuals' life happiness. The analyses were performed with a laptop with Intel i-7 2.3 GHz CPU and 8GB RAM.

5.2 Experimental results

5.2.1 Experimental result of classification analysis. In this experiment, we used Classification Analysis to analyze the experimental data set. The objective class we choose is the social individuals' life happiness. To prove the effectiveness of both our attribute selection approach and classification algorithm, we perform classification analyses with four different scenarios. Scenario 1 is to perform attribute selection on the data set and use random forest algorithm for classification analysis. Scenario 2 is to perform attribute selection on the data set and use decision tree algorithm for classification analysis. Scenario 3 do not perform attribute selection on the data set and use random forest algorithm for classification analysis. Scenario 4 do not perform attribute selection on the data set and use decision tree algorithm for classification analysis. The experimental result is shown in Table I. The classification accuracy within the table can be calculated as follows:

$$\text{Accuracy} = \frac{\textit{Number of correct classifications}}{\textit{Total number of classifications}} \tag{21}$$

The result shows that performing filter-based attribute selection on the GSS data set can slightly improve the classification accuracy of classification analysis and the Random

Table I.
Classification performance of different classification scenarios

	Scenario 1 Attribute Selection + Random Forest (%)	Scenario 2 Attribute Selection + Decision Tree (%)	Scenario 3 Random Forest (%)	Scenario 4 Decision Tree (%)
Classification Accuracy	85.0%	76.2%	84.1%	71.1%

Forest algorithm we choose can achieve better classification accuracy than the classic Decision Tree algorithm. The discriminative patterns-based classification framework generate top-20 discriminative patterns from decision trees in the forest. We chose ten valuable rules which are formed as Table II.

Through interpreting these rules, we obtain several interesting patterns from DPclass. Here, we give three prominent patterns as knowledge instances.

(1) The acceptance of social equity (question a35) is a major factor in determining happiness. Even people who are often depressed (a17_0=1) are more likely to live in a happy life if they feel that society is fairer. Conversely, people who lack of social trust (a33_DK = 1) and feel that society is unfair (a35_C = 1) are more likely to live unhappily.

(2) Subjective health feelings (question a15) and the objective physical health (question a16) has huge impact on life happiness. For instance, Married urban residents (a69_UM = 0) who feel very healthy (a15_VH = 1) more likely to feel happy in non-farm work (a58_ENA). People who has poor health condition which affect their career (a16_S=1), have no spouse (a69_FMHS = 0), and received elementary school degree or less (a72_BPS=1) are more likely to live unhappily.

(3) For these urban residents who are often depressed and have lower incomes (a8b_cut_l = 1), living space is a major factor that determine if they life happy or not. That is, they are much more likely to live in a happy life, as long as their living space is huge(a11_2 = 1).

5.2.3 Experimental result of the prediction analysis. In this experiment, we evaluated the effectiveness of Prediction Analysis by exploring the relationship between individuals' incomes and other factors. After converting the enumerated attributes as discrete attributes through attribute pre-processing, the exploration is performed by setting the individual's annual income as the dependent variable and all the other attributes as independent variables. The prediction analysis with RuleFit achieved Explained Variance Score as 0.52. The visualization of prediction results is shown in Figure 4, where red points are true values and yellow points are predicted values.

Rank	Discriminative rules
1	(a15_VU = 1) AND (a35_MU = 1)
2	(a11_1 = 1) AND (a17_F = 1) AND (a18_AA = 0) AND (a431_4 = 0)
3	(a13_3 = 1) AND (a431_1 = 1) AND (a33_DK = 0) AND (a59a_ENA = 0) AND (a64_FBA = 0)
4	(a431_3 = 0) AND (a50_B = 0) AND (a312_F = 1) AND (a59a_ENA = 1) AND (a64_BA = 0) AND (a64_A = 1)
5	(a13_1 = 0) AND (a14_4 = 1) AND (a51_G = 0) AND (a312_R = 1) AND (a35_DK = 0)
6	(a17_O = 1) AND (a52_CS = 0) AND (a312_O < 0.5) AND (a35_MU = 1)
7	(a15_G = 0) AND (a17_R = 0) AND (a431_1 = 1) AND (a69_FMHS = 0) AND (a16_S = 1) AND (a72_BPS) = 1)
8	(a15_VH = 1) AND (a69_UM = 0) AND (a16_rare = 0) AND (a58_ENA = 1) AND (a89b_subordinate = 0)
9	(a17_S = 1) AND (a18_A = 0) AND (a52_B = 0) AND (a33_DK = 1) AND (a35_C = 1)
10	(a18_AA = 0) AND (a431_3 = 0) AND (a431 < 3) AND (a16_N = 1) AND (a64_BA = 1)

Table II. Top ten discriminative rules generated by classification analysis

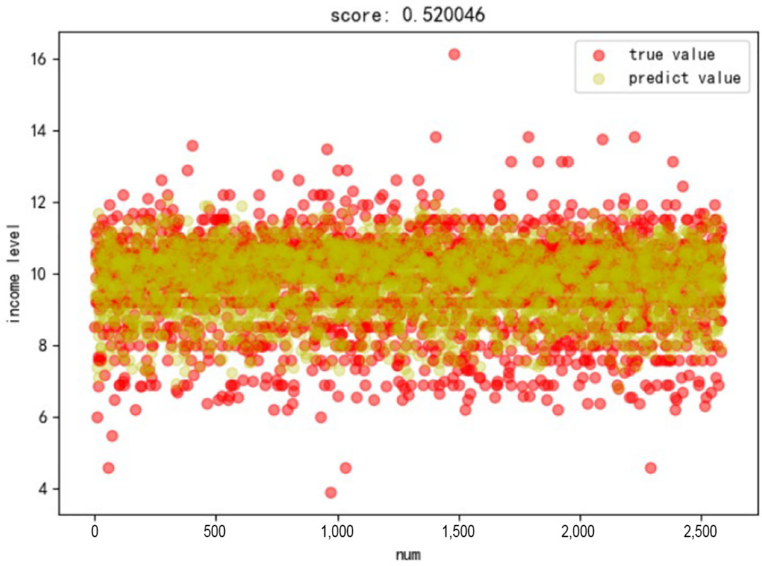


Figure 4.
The visualization of annual income prediction result

The interpretable patterns generated by RuleFit algorithm include both linear coefficient index and linear coefficient rules. 8 rules are listed in Table III as instances. The sample knowledges derived from these patterns are as follows:

- Among all the attributes, the correlation between individual's education level and annual incomes is the most obvious. Such correlation is described in rule 1.
- Rule 2 indicates that the population who received at least middle school education and live in city are likely to have higher income.
- Rule 3 indicates that the population who received at least bachelor degree and whose age is more than 27 are likely to have higher income.
- Rule 4 indicates that the male population whose age is more than 27 are likely to have higher income.
- Rule 5 indicates that population who think their social class are above average and whose age is more than 25 are likely to have higher income.

5.2.4 Experimental result of the clustering analysis. In this experiment, we evaluate the performance of clustering analysis. In order to improve the accuracy and interpretability of clustering analysis, we utilize the attribute selection to rank all the attributes based on their

Table III.
Sample coefficient rules generated by prediction analysis

NUM	Rules	Type	Coef	Support
1	edu_level	linear	0.01823	1
2	edu_level> 14.0 and age> 26.5	rule	0.20585	0.93805
3	edu_level> 3.0 and city> 0.5	rule	0.17240	0.57699
4	gender = male and age > 27.5	rule	0.07126	0.49380531
5	Social class> 5.5 and age> 24.5	rule	0.13586	0.185841

relevancy to the attribute named "degree of individual's life happiness". Among these highly relevant attributes, the attributes we chose are listed in Table III. The selection is made by taking both relevancy and experts' interest into consideration.

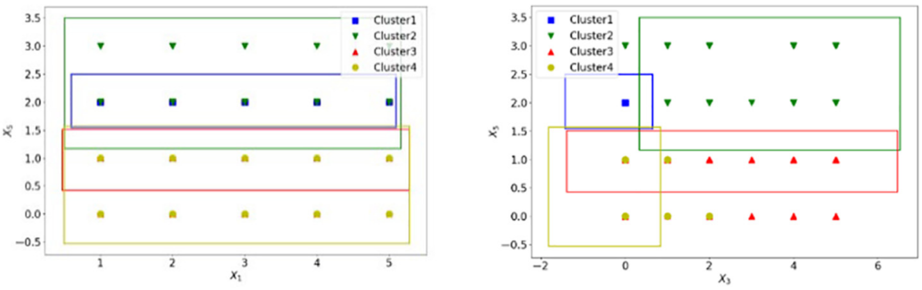
The clustering result generated by DReaM algorithm are four clusters whose boundary can be described by a set of rules shown in Table IV. To make more intuitive display, we can select a combination two attributes as the *x*-axis and *y*-axis to illustrate the distribution of their attributes values (see Figure 5), where the rectangles represent the boundary of clusters discovered by DReaM algorithm. As shown in Figure 5(a), the income level of people in Cluster 1 and Cluster 2 is higher, but the income level of people in Cluster 2 is wider than that of Cluster 1. The population represented by Cluster 3 is characterized by the general income level, and the population represented by Cluster 4 has the lowest income level. Since the boundary of clusters are roughly consistent with the attribute value distribution of "income level", we can conclude that the attribute "income level" is a decisive factor in this clustering analysis. Moreover, we can conclude that the correlation between income and life happiness are weak, and such conclusion is consistent with the result generated by classification analysis.

Figure 2(b) illustrate the cluster boundaries in the perspectives of "incomes level" and "frequency of Internet use". We can see that the data samples within Cluster 2 and 4 distribute tightly in this 2-D data space. Consequently, by interpreting the boundary of cluster 2, we can conclude that there is a portion of Chinese population who have the following two characteristics: obtains upper-middle-level incomes; and never surf the internet. By interpreting the boundary of cluster 4, we can conclude that there is a portion of Chinese population who have the following two characteristics: obtains lower-level incomes; and never surf the internet. Moreover, according to the boundary

Cluster1	Cluster2	Cluster3	Cluster4
$0.46 < x1 < 5.29$	$0.57 < x1 < 5.15$	$0.52 < x1 < 5.26$	$0.56 < x1 < 5.07$
$-3.02 < x2 < 17.33$	$-1.15 < x2 < 18.22$	$-0.55 < x2 < 18.37$	$-1.40 < x2 < 18.66$
$-1.70 < x3 < 0.74$	$0.65 < x3 < 6.39$	$0.41 < x3 < 6.46$	$-1.19 < x3 < 6.39$
$-1.25 < x4 < 2.60$	$-0.97 < x4 < 1.94$	$-0.88 < x4 < 5.67$	$2.08 < x4 < 5.77$
$-0.50 < x5 < 3.26$	$1.24 < x5 < 3.43$	$0.43 < x5 < 1.50$	$1.22 < x5 < 3.42$

Note: x1: Degree of happiness; x2: Years of received education; x3: Frequency of Internet use; x4: Learning frequency in idle time; x5: Income level

Table IV. Clusters generated by clustering analysis



(a) **(b)**

Figure 5. The visualization of clustering analysis results

of cluster 1, most of population who have higher-than-average income surf the internet during their daily life.

5.2.5 Experimental result of the association analysis. In this experiment, we used Association Analysis to analyze the experimental data set. We choose to analyze relationship between the social individuals' life happiness and other factors. The consequent of generated association rules is "happy = 1", which indicates the individuals live in a happy life. To prove effectiveness of association analysis, we evaluate the quality of association rules it generates through analyzing the experimental data set. By setting the minimum support level as 0.0005 and minimum confidence as 0.75, 796 association rules were generated within 5 seconds. Through filtering these rules based on their support level, confidence level, and knowledge theoretical value, 5 prominent rules are listed in Table V as instances. These rules can be interpreted as follows:

- The acceptance of social equality is a major factor for life happiness. For instance, according to rule 1, 89.98 per cent of the population who speak fluent mandarin and think the current social equality is acceptable live in a happy life. Moreover, according to rule 2, 89.97 per cent of the population whose living space is in average level (80-110 square meters) and think the current social equality is acceptable live in a happy life.
- Good health condition is a major factor for life happiness. For instance, according to rule 3, 90.46 per cent of the population who are in good health condition and think the current social equality is acceptable think they live in a happy life. Moreover, according to rule 4, 88.08 per cent people who are in good health condition and think they can trust most of the people in society think they live in a happy life.
- The key to life happiness for middle-class population are much more easier to found than population in other social class. For instance, according to rule 5, 88.67% of the middle-class population who are in good health condition think they live in a happy life. According to rule 6, 90.99 per cent of the middle-class population who are in marriage and have no divorce experience think they live in a happy life.

6. Conclusions

In this paper, we propose a comprehensive data management and analytic approach for General Society Survey data set based on a set of data mining algorithms. The approach can explore hidden patterns from the data set, which can be interpreted as knowledges for data

Table V.
Rules generated by association analysis

Antecedent	Consequent	Support	Confidence
mandarin = 4, social equality = 4	happy = 1	0.14732428	0.8997773
Living space = 4, social equality = 4	happy = 1	0.09162184	0.8997314
health = 1, social equality = 4	happy = 1	0.1667426	0.9045500
health = 0, social_trust = 4	happy = 1	0.1508798	0.8807877
health = 0, social class = 6	happy = 1	0.10839639	0.8866518
marriage = 3, social class = 6	happy = 1	0.08378157	0.9099010
rela = 4, stud = 3	happy = 1	0.05770809	0.8865546

driven policy-making purpose. To best of our knowledge, our work is the first work which proposed a data-mining based comprehensive data analytic approach for Society Survey data set. By implementing our approach, constructive knowledges regarding the key to social individuals' life happiness were extracted from the CGSS data set. It can be seen from results that the knowledges generated by different types of analysis mutually authenticate each other. As a future work, we plan to further improve the performance of our CBDMM approach by adding knowledge visualization function and integrating more data mining algorithms into it.

References

Australian Bureau of Statistics (2014), "1200.0.55.006 – age standard", available at: www.abs.gov.au/ausstats/abs@.nsf/Lookup/1200.0.55.006main+features62014,%20Version%201.7

Borgelt, C. (2005), "An implementation of the FP-growth algorithm", *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, ACM, pp. 1-5.

Davis, J.A. and Smith, T.W. (1991), *"The NORC General Social Survey: A User's Guide*", SAGE publications.

Dittman, D.J., Khoshgoftaar, T.M., Wald, R. and Napolitano, A. (2013), "Classification performance of rank aggregation techniques for ensemble gene selection", *The Twenty-Sixth International FLAIRS Conference*.

Du, P. and Yang, H. (2010), "China's population ageing and active ageing", *China Journal of Social Work*, Vol. 3 Nos 2/3, pp. 139-152.

Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. (2001), "Rank aggregation methods for the web", *Proceedings of the 10th international conference on World Wide Web*, ACM, pp. 613-622.

Friedman, J.H. and Popescu, B.E. (2008), "Predictive learning via rule ensembles", *The Annals of Applied Statistics. JSTOR*, Vol. 2 No. 3, pp. 916-954.

Gao, J., Liu, N., Lawley, M. and Hu, X. (2017), "An interpretable classification framework for information extraction from online healthcare forums", *Journal of Healthcare Engineering*, Vol. 2017, doi: 10.1155/2017/2460174.

Hu, A. and Leamaster, R.J. (2015), "Intergenerational religious mobility in contemporary China", *Journal for the Scientific Study of Religion*, Vol. 54 No. 1, pp. 79-99.

Johnston, M.P. (2017), "Secondary data analysis: a method of which the time has come", *Qualitative and Quantitative Methods in Libraries*, Vol. 3 No. 3, pp. 619-626.

Kruidenier, L.M., Nicolaï, S.P.A., Willigendael, E.M., *et al.* (2009), "Functional claudication distance: a reliable and valid measurement to assess functional limitation in patients with intermittent claudication", *BMC Cardiovascular Disorders*, Vol. 9 No. 1, p. 9.

Lorenzo, R. (2013), "*Individual Income Tax Law, Chinese Tax Law and International Treaties*, Springer International Publishing, pp. 9-21.

Mitra, P., Murthy, C.A. and Pal, S. (2002), "Unsupervised feature selection using feature similarity", *IEEE Trans. Pattern Anal. Mach. Intell*, Vol. 24 No. 3, pp. 301-312.

National Survey Research Center (NSRC) at Renmin University of China (2019), "Chinese General Society Survey, 2019", available at: http://cgss.ruc.edu.cn/index.php?r=index/index&hl=en

Statistics Canada (2017), "Age categories, life cycle groupings", available at: www.statcan.gc.ca/eng/concepts/definitions/age2

Tan, H. (2014), "The problems in rural English teaching and the optimization path: a study based on the Chinese general social survey data", *Asian Agricultural Research*, Vol. 6 No. 1812-2016-143451, pp. 86-92.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58 No. 1, pp. 267-288.

Wu, X., Ye, H. and He, G.G. (2014), "Fertility decline and women's status improvement in China", *Chinese Sociological Review*, Vol. 46 No. 3, pp. 3-25.

Zhao, Z. and Liu, H. (2007), "Spectral feature selection for supervised and unsupervised learning", *Proceedings of the 24th international conference on Machine learning*, ACM, pp. 1151-1157.

Corresponding author
Zhiwen Pan can be contacted at: pzw@ict.ac.cn