Anomaly data management and big data analytics: an application on disability datasets

Zhiwen Pan, Wen Ji and Yiqiang Chen
Institute of Computing Technology Chinese Academy of Sciences, Beijing, China, and
Lianjun Dai and Jun Zhang
Information Centre of China Disabled Persons' Federation, Beijing, China

Received 8 September 2018
Revised 12 October 2018
Accepted 13 October 2018

Abstract

Purpose – The disability datasets are the datasets that contain the information of disabled populations. By analyzing these datasets, professionals who work with disabled populations can have a better understanding of the inherent characteristics of the disabled populations, so that working plans and policies, which can effectively help the disabled populations, can be made accordingly.

Design/methodology/approach – In this paper, the authors proposed a big data management and analytic approach for disability datasets.

Findings – By using a set of data mining algorithms, the proposed approach can provide the following services. The data management scheme in the approach can improve the quality of disability data by estimating miss attribute values and detecting anomaly and low-quality data instances. The data mining scheme in the approach can explore useful patterns which reflect the correlation, association and interactional between the disability data attributes. Experiments based on real-world dataset are conducted at the end to prove the effectiveness of the approach.

Originality/value – The proposed approach can enable data-driven decision-making for professionals who work with disabled populations.

Keywords Decision support systems, Data analytics, anomaly data detection, Data management systems

Paper type Research paper

1. Introduction

With the development of the big data management and storage techniques, more and more data which contains the information of disabled population has been collected. Through analyzing these precious disability datasets, people can gain knowledges such as what are the living conditions and demands of the disabled population and how the current assisting services work for them (Mcdermott and Turk, 2015). In this way, the professionals who work with disabled

This work is supported by the National Key R&D Program of China (2017YFB1400100), the National Natural Science Foundation of China (61572466), the Beijing Natural Science Foundation (4162059), and the research on "Big Data Analysis and Data Mining for Disabled Persons".

International Journal of Crowd Science
Vol. 2 No. 2, 2018
pp. 164-176
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-09-2018-0020

populations can have a better understanding of how to make working plans and policies to help the population in a right way (Janssen *et al.*, 2017). However, researchers have pointed out that there are two major challenges for managing and analyzing the disability datasets with traditional methods (Hoffman, 2017; Nambisan *et al.*, 2015). First, as most of the data are collected manually from the disabled individuals, the data qualities (e.g. data accuracy, data authenticity, missing values, etc.) are highly affected by the individual's subjective factors. Second, as disability datasets usually consist of a huge amount of disability data attributes which are correlated and interactional with each other, traditional data statistics methods are not intelligent enough to explore these correlation and interaction between the data attributes. Hence, to better support the disabled populations. An intelligent approach which can improve the quality of data and perform comprehensive and effective analysis on disability datasets is needed.

By leveraging the power of data mining techniques to the analytics of disability datasets, we proposed a Big Data Management and Mining (BDMM) approach which can provided the professions a unified and effective way to manage and analyze the disability datasets. To meet the aforementioned two challenges, our BDMM approach is designed to provide two services through two working phases (Figure 1). During the data management phase, interpolation analysis and anomaly detection analysis are performed to improve the quality of disabled datasets by estimating missing attribute values and detecting anomaly and low-quality data instances. During the data mining phase, useful patterns which describe the correlation, association and interactional between the disability data attributes can be fully explored through performing all the major types of data mining analytics which include association analysis, classification analysis and prediction analysis (Fuji and Matsumoto, 2018; LIU, 2014; Akyildiz *et al.*, 2002). These patterns can be used to assist the professions making decisions and policies, in this way, the data-driven decision-making can be performed.

The major contributions of this paper are as follows:

- We proposed the first comprehensive big data management and mining approach for Sociological datasets of disability population, and it is possible to apply this approach on other kinds of sociological datasets.

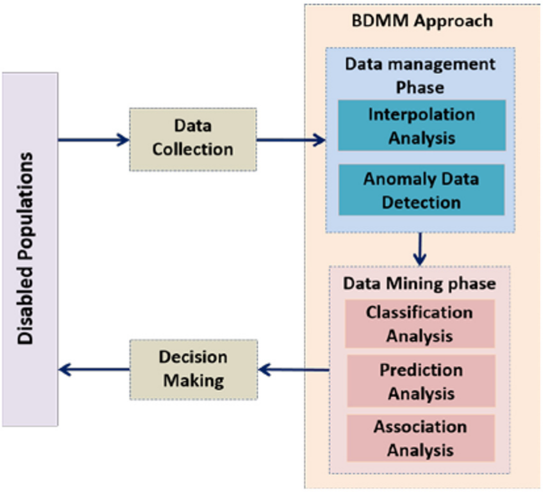

Figure 1.
The flow diagram of Big Data management and mining approach

- In the data management phase of our approach, we proposed an anomaly detection scheme which combines data dimension reduction algorithm with density based anomaly detection algorithm.
- To help profession performing data-driven decision making, in the data mining phase of our approach, we used a set of data mining algorithms which can generate interpretable patterns.

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

2. Big data management and mining approach: data management phase

2.1 Data interpolation analysis

The Data Interpolation Analysis is to fill in the missing attribute values within disabled dataset with estimated values. The estimation of missing values is made by performing interpolation operation on disability datasets. The algorithm we choose to perform data interpolation analysis is Cubic Spline Interpolation (CSI). We choose this algorithm since it can achieve better estimation accuracy than other interpolation algorithms such as linear interpolation and piecewise interpolation (Mckinley and Levine, 2007). CSI generates locally fitting function (which is a cubic function) based on data within each local section. As the generated fitting function is a curve that is piecewise connected, its gradient is usually smoother than the ones generated by other interpolation algorithms. The fitting function generated by CSI is represented as:

$$\begin{cases} C_1(x), x_0 \leq x \leq x_1 \\ \dots \\ C_n(x), x_{n-1} \leq x \leq x_n \end{cases} \tag{1}$$

Where each C_i is a cubic function as:

$$C_i = a_i + b_i x + c_i x^2 + d_i x^3 \tag{2}$$

For each local section $[x_i, x_{i+1}]$, we define its step size h_i as $h_i = x_{i+1} + x_i$. The four parameters in equation (2) can be calculated as:

$$\begin{cases} a_i = y_i \\ c_i = \frac{1}{2} S_i^n(x_i) = m_i \\ b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{2} m_i - \frac{h_i}{6}(m_{i+1} - m_i) \\ d_i = \frac{m_{i+1} - m_i}{6h_i} \end{cases} \tag{3}$$

By generate the fitting function S(x) for a dataset, we can perform interpolation analysis to estimate values of the dataset within and beyond the section of $[x_0, x_n]$. For instance, given the disability income dataset that contain disabled individual's incomes between 2008 and 2018, if the dataset contains missing attribute values in the year of 2012, we can estimate the missing values based on the fitting function.

2.2 Anomaly detection analysis

As the collection of disability data is through letting subjects fill in forms and questionnaires, it is possible that some of the subjects will provide fake or inaccurate information for a variety of reasons. Moreover, some of the data may be stored into the database incorrectly because of technical fault. When being used for analysis, these fake and inaccurate data may cause drift of the analytic model, hence resulting in inaccurate analytic results. To dynamically detect and remove these fake and inaccurate data, Anomaly Detection Analysis generates a baseline model which can describe the patterns of correct data so that any data whose pattern is different from the normal pattern will be detected as the suspicious data. The operation of anomaly detection analysis is based on the assumption that a majority of the data within disability datasets are correct. Hence, data that are extremely different from most of the other data will be regarded as fake or inaccurate data.

The analysis uses a distance-based anomaly detection algorithm named local outlier factor (LOF) (Ma *et al.*, 2016). The three reasons for choosing this algorithm are as follows:

1. It is an unsupervised algorithm which can perform data training based on unlabeled data, which means we do not know to have priori knowledge about the anomaly data.
2. It is a density-based algorithm which has relatively lower computational complexity and is capable of describing the pattern of data that are distributed in all kind of shapes (circle, bar, ring, etc.).
3. It can quantify the abnormality of each data by calculating its LOF.

To improve the performance (both detection speed and detection accuracy) of our anomaly detection analysis, we utilize a dimension reduction algorithm named principle component analysis (PCA), which can merge the relevant attributes as a one-dimension attribute named principle factor (Wold *et al.*, 1987). Based on expert knowledges, we manually classify all the disability datasets attributes into these five categories, and then use PCA to merge high-dimensional disability datasets attribute-set into a five-dimensional matrix which contains the five principle factors. Given a dataset matrix denoted as $X_{n\times m}$ and an output which is a decision matrix denoted as $D_{1\times m}$, the pseudo-code of our anomaly detection algorithm is as follows:

```
AnomalyDetection($X_{n\times m}$):

  use Principle Component Analysis algorithm to convert $X_{n\times m}$ to factor matrix $X'_{5\times m}$

  for each data sample $x'_{5\times 1}$ and $o'_{5\times 1}$ in $X'_{5\times m}$ do

  calculate the Euclidean distance $d(x', o')$ of $x'_{5\times 1}$ with all the other data samples $o'_{5\times 1}$
```

```
    locate the K-th nearest point $N_k(x')$ of $x'_{5\times1}$ and set the distance between the two points as $k\text{-}dist(x', o)$ of $x'_{5\times1}$
    calculate the reachable distance $reach\text{-}dist(x', o')$ between $x'_{5\times1}$ and points $o'$ based on equation: $reach\text{-}dist(x', o') = \max(k\text{-}dist(x'), d(x', o'))$
    calculate the local outlier density $\mathrm{lrd}(x')$ based on equation $\mathrm{lrd}(x') = \frac{1}{\frac{\sum_{0\in N_k(x')}\frac{lrd(o')}{lrd(x')}}{|N_k(x')|}}$
    calculate die LOF $\mathrm{lof}(x')$ based on equation $\mathrm{lrd}(x') = \frac{\sum_{0\in N_k(x')}\frac{lrd(o')}{lrd(x')}}{|N_k(x')|}$
    if $\mathrm{lof}(x') > threshold$ then
        $D(x') \leftarrow anomaly$
    end if
end for
return $D_{1\times m}$
```

3. Big data management and mining approach: data analytics phase

3.1 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. For instance, if a disability dataset consists of attributes such as Employment Condition (EC), Education Level (EL) and Handicap Category (DC). If we find that the attribute value DC = “hearing handicap” and EL = “middle-school” always appear together with EC = “no job”, we can interpret this association as: “It is very hard for the auditory handicapped subjects who only finish middle school to find a job.” By performing association analysis on disability datasets, we can generate a huge set of association rules. These rules can give practitioner a better understanding of the Chinese disabled population. Hence, the practitioners can make plans and policies accordingly to serve the disabled population better.

The association analysis performs association analysis by utilizing a machine learning algorithm named FP-growth (Borgelt, 2005). We choose this algorithm since its computational complexity is low than other algorithms such as Apriori. We implement FP-growth algorithm based on the following two steps. In the first step, a tree graph is built to describe data distribution of the training dataset. Given a dataset matrix $T_{n\times m}$, the support level of each attribute value x_i within the dataset is calculated based on the following equation:

$$\mathrm{supp}(x_1) = \frac{|\{t \in T; x \in t\}|}{|T|} \tag{4}$$

where t is a row within matrix $X_{n\times m}$ and $\left|\{t \in T; x \in t\}\right|$ is the number of rows which contain the attribute value x. Then the analysis selects the attribute values whose support levels are above a pre-defined threshold as frequent items. In the second step, Frequent Pattern tree (FP-tree) is built based on frequent items so that association rules can be extracted. Starting from the empty set $\varnothing$, the analysis stores the attribute value $x_1$with the highest support level in the top layer of FP-tree. Then, the analysis finds all the data instances which appear together with x_1 and store them in the second layer of FP-tree. Such process is repeated in a recurrent manner until all the attribute values are stored into the FP-tree. A sample FP-tree is shown in Figure 2. The label in each block indicates the name of attribute value and its number of occurrence in the training dataset. Associate rules are extracted from the FP-tree by determining if the confidence level of a rule is above a predefined threshold. The confidence level is calculated based on the following equation:

$$conf(X \Longrightarrow Y) = supp(X \cup Y)/supp(X) \tag{5}$$

where X and Y are frequent item sets (e.g. an association rule $X \Rightarrow Y$ can be $\{x_1, x_5\}\{x_3\}$).

One of the major drawback of association analysis is that it is sensitive to data imbalance problem which means that the frequent attribute values will prevent the algorithms from analyzing the less frequent attributes (Longadge and Dongre, 2013). Such problem becomes even worse when analyzing a big dataset such as the Chinese Disability Dataset. To solve this problem, we made the following two efforts. First, we set the value of minimum support level and minimum confidence level as extremely low. In this way, the lesser frequent attribute values can also be analyzed. However, since minimum support level and minimum confidence level will make the number of rules been generated grow exponentially, the second effort we made is to set the maximum length of rules that should be generated as either two or three (both can be regarded as short-length), which means that only one or two antecedent and one consequent is allowed. In this way, more than 90 per cent of the rules will be filtered out. By analyzing these short-length rules, we can determine the important antecedents and then filter the long-length rules with these important antecedents.

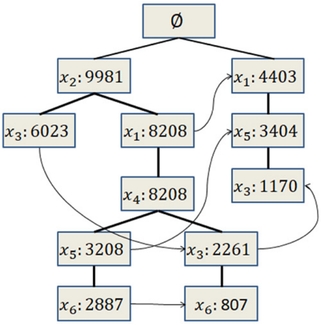

Figure 2.
A sample FP-tree generated by association analysis

3.2 Classification analysis

The disabled population can be classified into multiple categories for different purposes. For instance, according to their incomes, the disabled population can be classified into category such as population with extremely low income, population with low income and population with medium or high income. Based on these pre-defined population categories, the Classification Analysis aims at exploring the unique characteristics of different populations. For instance, through performing classification analysis to classify the categories of income, the analysis may find a set of characteristics (or patterns) such as:

- most of the people whose disability is extremity disability and education level is elementary school or lower are earning extremely low income; and
- most of people who is married and lives in a city are earning medium or high income.

These patterns can be used to determine the major differences between populations in different categories.

In our classification analysis, we choose to apply Decision Tree algorithm which is a classic classification algorithm (Salzberg, 1994). We choose this algorithm for two reasons. First, this algorithm constantly achieves better classification accuracy than other algorithms on disability datasets. Second, the tree model generated by Decision Tree algorithm is interpretable, this means that the domain experts can interpret the patterns described by tree model into knowledges that can be easily understood. During the data training process, the Decision Tree algorithm generates tree model in a top-down manner by following a predefined metrics such as Gini impurity. Specifically, the algorithm tries to split the data samples into subsets so that categories of data samples within each subset can be classified separately with better classification accuracy. During each iteration of split, the split which generates subsets with lowest Gini Impurity will be selected. The Gini impurity is a light-weight version of the famous metrics named Information Gain. The Gini impurity of a subset can be calculated as:

$$\mathrm{I}(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1 - p_i) = 1 - \sum_{i=1}^{J} p_i^2 \tag{6}$$

where p_i is the fraction of items labeled with class i in the subset. Such split is performed in a recursive manner which means the classification model generated by Decision Tree algorithm is a tree model with multiple layers where each layer contains a set of subsets.

3.3 Classification analysis

The prediction analysis aims at evaluating the influence of some attributes to the variation of a certain numerical attributes within disability datasets. 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. For instance, if we want to determine what attributes may affect the annual income of disabled population, we can predict the variation trend of annual income based on a set of candidate attributes. If the prediction result shows that the weighted factors of attributes including annual expense, employment condition and handicapped level are bigger than that of other attributes, we can conclude that these three attributes have effects on the income of the disabled population. If the weight factor of an attribute is positive, the effect is positive; otherwise the effect is either zero or negative.

The prediction analysis performs the prediction by utilizing the linear regression algorithm (Olive, 2013). This algorithm use linear function to model the relationship between a scalar dependent variable $y = (y_1, y_2, \ldots, y_n)$ and a set of explanatory variables $\mathrm{v} = \left(x_1^T, x_2^T, \ldots, x_n^T\right)$, where T denotes the transpose and x_n^T is a $1 \times m$ matrix representing a variable. The linear regression algorithm assumes that the relationship between the two kinds of variables is linear, and such relationship takes the form as:

$$y_i = \theta_0 1 + \theta_0 x_{i1} + \ldots + \theta_0 x_{in} + \varepsilon_i \tag{7}$$

where ε_i is a disturbance term for the i-th value of variables. By stacking the equation (6) as vector form, the relationship can be represented as:

$$Y = h_\theta(X) = \theta^T X + \varepsilon \tag{8}$$

To optimize the quality of $h_\theta(X)$, the analysis adjust the value of θ^T to minimize an objective function:

$$\mathrm{J}\left(\theta^T\right) = \min\left(\frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x_n) - y_i\right)^2\right) \tag{9}$$

Through utilizing the least square approach, the analysis makes adjustment based on the following equation:

$$\theta = \left(X^T X\right)^{-1} X^T Y \tag{10}$$

In this way, the values within matrix θ are determined.

4. Experimental evaluation and results

4.1 Experimental dataset

To prove the effectiveness of our proposed approach, we implemented the approach to analysis a dataset which consists of 33 million data instances and it is stored as an 8GB csv file. These data instances are collected from 33 million Chinese disabled individuals through questionnaires. There are around 50 attributes within the dataset which describe the disabled population in four perspectives: the disabled individual's personal information, the individual's living condition, assisting services been received and the current demand. Restricted by the privacy policies of this dataset, we cannot reveal more details of the dataset. Before performing the anomaly detection and data analytics, the data preprocessing works we did are as follows:

- Since some of the attributes are answers of the multiple-choice questions, we decompose such attribute into a set of attributes which have one-to-one correspondence to the choices. For instance, if an attribute is the answer of a multiple-choice question with five choices, we decompose this attribute into five binary attributes indicating which choices are included in the current answer.
- With the help of domain experts, we remove some attributes are useless for performing data analytics from the dataset.
- We used data interpolation analysis to fill in the missing attribute values.

4.2 Experimental results

4.2.1 Performance evaluation of the anomaly data detection. In this experiment, we evaluate the performance of the proposed Anomaly Data Detection analysis. By performing data dimension reduction with PCA algorithm, the analysis merge the attributes within the experimental dataset into four attributes which include the employment and income factor, personal information factor, medical and health factor, and community service factor. After performing anomaly detection on this new dataset, 30 thousands anomaly data instances were detected from the 3300 thousands data instances. The visualization of our anomaly detection result is shown in Figure 3, where the blue and red points represent the normal and anomaly data instance respectively. When performing classification analysis on dataset without these anomaly data instances, the classification accuracy is increased from 67.0 to 68.71 per cent.

4.2.2 Experimental result of classification analysis. In this experiment, we used classification analysis to analyze experimental dataset. The classification analysis was performed by classifying the disabled population based on their genders so that the difference between male and female disability population can explored. Through performing data training with Decision Tree algorithm, tree model with 391 nodes been generated. A visualization of the decision tree model is shown in Figure 4. As shown in the figure, the information within each node include the splitting condition, current gini impurity, number of data samples covered by the split, the class distribution of the data samples. The color shade of nodes indicate the quality of the last split (darker color indicates better quality). The classification accuracy achieved by the tree model is 68.71 per cent. This classification rate is acceptable since our analysis on disability dataset can be regarded as sociological analysis, and it is not enough to fully describe the disability population with our current dataset. Through interpreting these rules, a lot of interesting patterns have been found. Here, we take three of these patterns as instances:

- The population who are not married and whose ages are older than 29 consist much more males than females (ratio: 3.44:1).

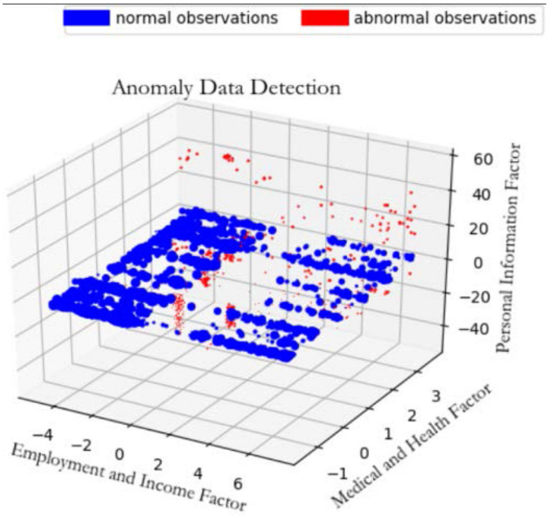

Figure 3.
A visualization of anomaly data detection results

- The population who are married and received at least middle school education consist of much more males than females (ratio: 2.09:1).
- The rural population whose age are older than 29, not married consist of much more males than females (ratio: 6.52:1).
- The rural population whose age are between 29 and 39, not married, not receive high school education, consist of much more male than female (ratio: 13.63:1).
- The population who are widowed consist of more female than male (ratio: 4.85:1).
- The population who received no education consist of more female than male (ratio: 1.73:1).
- The population who are married received no education consist of more female than male (ratio: 2.27:1).

Based on these patterns, we can get the following conclusions:

- Marriage and education level are two of the major factors that differentiate male and female disability populations.
- The male disability population are facing bigger problem of finding a spouse. Such problem is more severe for the rural male disability population. Moreover, such problem is more severe for the rural male disability population who did not receive high school education.
- The female disability population are much more likely to be widowed.
- The overall education level of female disability population are lower than male disability population. This phenomenon is majorly because that there are much more female population who receive no education.
- Compared to male disability population, the factor of low education does not obviously prevent female disability population from finding a spouse.

4.2.3 Experimental result of the prediction analysis. In this experiment, we aim at evaluating the effectiveness of Prediction Analysis. The dependent variables we predicted are also the gender of disabled individuals (male and female). After performing data training with a dataset including 3,300 million data samples and 48 explanatory variables, a regression function that describe the difference between male and female disability population is generated. A visualization of that describe some of the details of the regression function is shown in Figure 5, where each cylinder describe

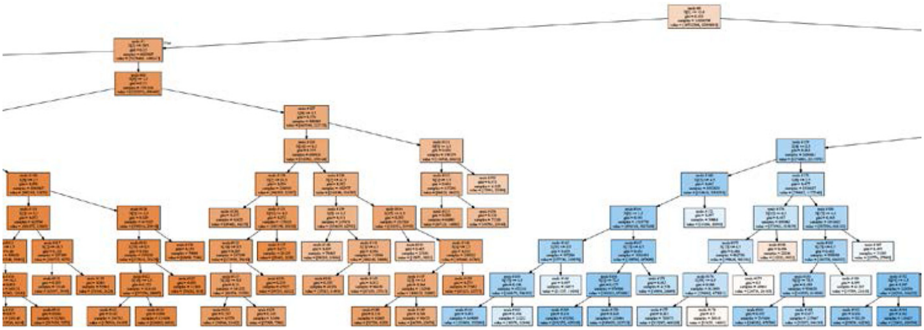

Figure 4.
A visualization of classification analysis results

the influential weight of an explanatory variable when determine if a disability individual is female or not, and names of the explanatory variables are shown in the horizontal axis. Note that all the influential weights were normalized so that they are comparable with each other. Based on the figure, the information we can infer are as follows:

- The major factors that determine if a disability individual is male or females include education level, marriage status, if at school or not, if at special school or not, if being employed or not, and if live in urban.
- There are more female disabled individuals who live in urban area.
- There are more female disabled individuals who are married.
- There are more female disabled individuals who are currently study at either common school or school for disability children.
- Female disabled individuals received slight more social welfare and disability assistance than male disability individuals.
- Female disabled individuals received slight more social welfare and disability assistance than male disability individuals.
- Male disabled population received more educations than female disability population.
- Female disabled individuals are more likely to be employed than male disability individuals, although male individuals received slightly more employment assistance.
- Compared with female disabled individuals, slightly more male disabled individuals feel that they need assistances for finding a job.
- Male disabled individuals received slight more social assistance than female disability individuals.
- Other than the disabilities, male disabled individuals are more likely to have other diseases.

4.2.4 Experimental result of the association analysis. To prove effectiveness of association analysis, we evaluate the quality of association rules it generates through analyzing the experimental dataset. As performing association analysis on our disability dataset could generate a huge amount of rules. In this experiment, we choose to generate the rules whose consequent is "Gender=Female". The experimental dataset we use is a dataset with unbalanced data since it has some attribute values that appear more frequently than others

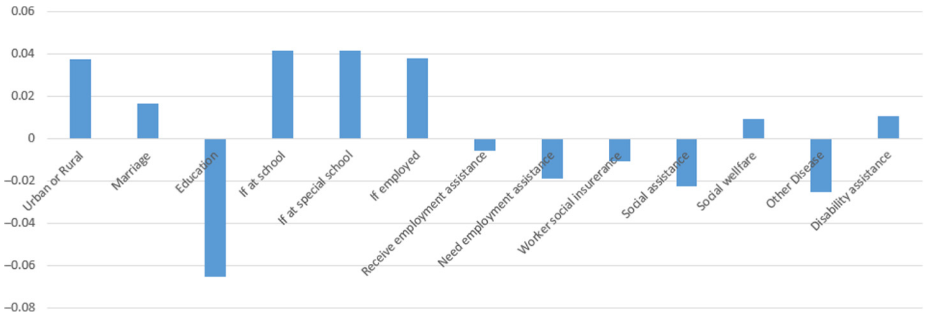

Figure 5.
A visualization of prediction analysis results

do. When performing association analysis on our experimental dataset, we encounter a severe data imbalance problem which means that the frequent attribute values will prevent the algorithms from analyzing the less frequent attributes (Longadge and Dongre, 2013). To solve this problem, we made the following efforts:

- We set the value of minimum support level and minimum confidence level as 0.00005 and minimum confidence as 0.1 which are both extremely.
- We set the maximum length of rules as two which means only one antecedent and one consequent is allowed. After the training process, 167 association rules whose consequent are "Gender = Female" are generated.

Some of these rules are listed as follows:

- {Disability category = extremity disability} => {Gender = female} support: 0.209, confidence: 0.3846745. This means that there are 20.9 per cent of the disability population are extremity disabled, and 38.47 per cent of them are females.
- {Studying at special school=junior college} => {Gender = female} confidence: 0.458. This means that, among the disability population who is pursuing junior college degree at colleges for disability population, there are 45.8 per cent of the females. Since there are 40.8 per cent of females in the disability population overall. This rule indicates that female disabled individuals are slightly more likely to be admitted by junior special colleges.
- {Living resource if have no job=retirement pension}=> {Gender=female} confidence: 0.6112235. This means that, among the disability population who used to have a job and currently retired, there are 61.12 per cent of the females. Since there are 40.8 per cent of females in the disability population overall. This rule indicates that compared with male disabled individuals, there are more female disabled individuals who are currently receiving retirement pension.

5. Conclusions

In this paper, we proposed a data management and analytic approach for disability data based on a set of data mining algorithms. The approach can improve the quality of disability big data, and then explore useful information from it. These information been explored can help professionals making policies and decisions to improve the wellbeing of disable population. The experimental results prove the effectiveness of our approach. Furthermore, since the results generated by different analysis are all regarding the difference between male and female disability population, the results generated by different analyses mutually proofed each other. For instance, all the three analyses have shown that:

(1) there are more female population who are at school than that of male population;

(2) the female population are more likely to be married than male population; and

(3) male population need more employment assistance than female population, etc.

As a future work, we plan to further improve the performance of our BDMM approach by adding the clustering analysis into it.

References

Akyildiz, I.F., Su, W., Sankarasubramaniam, Y. and Cayirci, E. (2002), "Wireless sensor networks: a survey", *Comm. ACM*, Vol. 38 No. 4, pp. 393-422.

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

Fuji, R. and Matsumoto, K. (2018), *Emotion Analysis on Social Big Data [J/OL]*, ZTECommunications. pp. 10-12, available at: http://kns.cnki.net/kcms/detail/34.1294.TN.20180105.1738.004.html

Hoffman, S. (2017), "Big data and the americans with disabilities act", *Social Science Electronic Publishing*, Vol. 68 No. 4, pp. 777-793, available at: www.acm.org/class/how_to_use.html

Janssen, M., van der Voort, H. and Wahyudi, A. (2017), "Factors influencing big data decision-making quality", *Journal of Business Research*, Vol. 70, pp. 338-345.

LIU, A. (2014), "Big data and its application in tourism industry: a theoretical review and analysis[a]", *Proceedings of the International Conference on Management and Engineering (CME 2014)*, 7.

Longadge, R. and Dongre, S. (2013), "Class imbalance problem in data mining review", arXiv preprint arXiv: 1305.1707.

Mcdermott, S. and Turk, M.A. (2015), "What are the implications of the big data paradigm shift for disability and health?[J]", *Disability and Health Journal*, Vol. 8 No. 3, pp. 303-304.

Mckinley, S. and Levine, M. (2007), "Cubic spline interpolation", *Methods of Shape-Preserving Spline Approximation*, pp. 37-59.

Ma, M.X., Ngan, H.Y.T. and Liu, W. (2016), "Density-based outlier detection by local outlier factor on largescale traffic data", *Electronic Imaging*, Vol. 2016 No. 14.

Nambisan, P., Luo, Z., Kapoor, A., Patrick, T.B. and Cisler, R.A. (2015), *Social Media, Big Data, and Public Health Informatics: Ruminating Behavior of Depression Revealed through Twitter[C]//HI International Conference on System Sciences. IEEE*, pp. 2906-2913.

Olive, D.J. (2013), "Linear regression analysis", *Technometrics*, Vol. 45 No. 4, pp. 362-363, available at: http://doi.acm.org/10.1145/503376.503378

Salzberg, S.L. (1994), *Machine Learning*, Vol. 16 No. 3, p. 235, available at: https://doi.org/10.1007/BF00993309

Wold, S., Esbensen, K. and Geladi, P. (1987), "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems*, Vol. 2 Nos 1/3, pp. 37-52.

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